

Sommersemester 2001

Interuniversitäres Seminar
Datenbanktechnologie für das Web

Universität Konstanz
Universität Zürich

Information Retrieval und das Web: Grundlagen & Problematik

Martin Waldburger
Gartenstrasse 13
CH-8807 Freienbach
wald@access.unizh.ch

Juni 2001

Dozent: Prof. Dr. K. R. Dittrich
Betreuung: Ruxandra Domenig

1.	Einführung	3
2.	Information Retrieval Grundlagen.....	4
2.1.	Begriff Information Retrieval.....	4
2.2.	Wie funktioniert ein Information Retrieval System	5
2.2.1.	Indexierung	5
2.2.2.	Retrieval Modelle.....	8
2.2.3.	Messung.....	11
3.	Information Retrieval im Web.....	12
3.1.	Spezialitäten des Web	12
3.1.1.	Datenmenge	12
3.1.2.	Dynamik	14
3.1.3.	Heterogenität der Daten.....	14
3.1.4.	Verschiedene Sprachen	15
3.1.5.	Heterogenität der Benutzer.....	15
3.1.6.	Duplikate	15
3.1.7.	Hohe Verlinkung.....	16
3.1.8.	Länge der Abfrageergebnisse und Benutzerverhalten	16
3.1.9.	Crawler, Roboter.....	18
3.2.	Arten von Information Retrieval Systemen für das Web.....	18
3.2.1.	Suchmaschinen	18
3.2.2.	Metasuchmaschinen	19
3.2.3.	Agenten.....	19
3.2.4.	Kataloge.....	19
4.	Beurteilung	20
5.	Literaturverzeichnis	21

1. Einführung

Die vorliegende Seminararbeit ist grundsätzlich zweiteilig aufgebaut. Der erste Teil befasst sich mit dem Thema Information Retrieval im allgemeinen und liefert die nötigen Grundlagen für den zweiten Teil. Dieser behandelt Information Retrieval im Web und geht auf dessen Spezialitäten und die dafür gefundenen Lösungen ein.

Wenn man das Problem des Information Retrieval historisch betrachtet, so lässt sich beobachten, dass es vor allem seit dem Ende des Zweiten Weltkriegs mit wachsendem Interesse verfolgt wird. Es kann relativ einfach umschrieben werden: Wir haben riesige Mengen an Informationen und präziser und schneller Zugriff darauf wird immer schwieriger. Ein Effekt davon ist, dass relevante Information nicht beachtet wird und zwar aus dem Grund, weil sie nicht mehr aufgefunden wird. Computer können uns zumindest behilflich sein, mittels passender Information Retrieval Systeme relevante, präzise und schnelle Ergebnisse auf unsere Suche nach Information zu geben.

Im Prinzip ist Information Retrieval nicht sonderlich schwierig: Stellt man sich einen Bücherladen vor und eine Person, die darin ein spezielles Buch sucht, so könnte diese Person ganz einfach alle Bücher durchblättern und so Relevantes von Irrelevantem trennen. Dieses Vorgehen der erschöpfenden Suche ist in der Praxis allerdings nicht praktikabel. Niemand hat die Zeit oder Lust, einen ganzen Bücherladen zu durchsuchen, nur um ein spezielles Buch zu finden.

Deshalb wird zur Lösung des Problems von Information Retrieval auf automatische Information Retrieval Systeme gesetzt. Das Aufkommen solcher Systeme wurde dadurch unterstützt, dass man mittlerweile über sehr leistungsfähige Computertechnologie zu verhältnismässig geringen Kosten zur Unterstützung des Prozesses verfügt. Natürlich kann auch ein noch so leistungsfähiger Computer von sich aus nicht entscheiden, was relevant und was weniger relevant ist – dafür müssen je nach Einsatzgebiet des Information Retrieval Systems verschiedene, angepasste Methoden und Algorithmen zum Einsatz kommen.

2. Information Retrieval Grundlagen

2.1. Begriff Information Retrieval

Eine genaue und eindeutige Definition des Begriffes Information Retrieval existiert leider nicht. Das ist wahrscheinlich dadurch begründet, dass das Gebiet sehr vielfältig und komplex ist. Der Begriff Information Retrieval wird zumeist in erweiterter Form unter "Information Storage and Retrieval" in den Enzyklopädien behandelt und folgendermassen definiert:

Information Storage and Retrieval, the systematic process of collecting and cataloging data so that they can be located and displayed on request¹.

Ins Deutsche übersetzt bezeichnet der Begriff also das systematische Vorgehen, um Daten zu sammeln und derart zu katalogisieren, dass sie auf Anfrage wieder aufgefunden und angezeigt werden können.

Aufgrund der sehr breiten Definition bietet sich eine Abgrenzung zum Datenretrieval an, um ein klareres Bild davon zu gewinnen, was Information Retrieval bedeutet. Rijsbergen² hat zu diesem Zweck eine Vergleichstabelle der beiden Retrieval-Arten zusammengestellt:

Vergleichskriterium	Datenretrieval	Information Retrieval
Matching	Exakt	Partiell, "best match"
Inferenz	Deduktion	Induktion
Modell	Deterministisch	Probalistisch
Anfragesprache	Formal	Natürlich
Fragespezifikation	Vollständig	Unvollständig
Gesuchte Objekte	Die Fragespezifikation erfüllende	Relevante
Reaktion auf Datenfehler	Sensitiv	Nicht sensitiv

Abbildung 1 - Vergleichstabelle Datenretrieval / Information Retrieval

¹ <http://www.encyclopedia.com/articles/06363.html>

² [VRIJS]

Daten- und Information Retrieval lassen sich ganz allgemein auch durch die Definitionen von Daten, Wissen und Information unterscheiden. Bei Daten ist lediglich die Syntax bekannt. Es kann sich beispielsweise um eine Zahlenfolge handeln. Bekommen die Daten eine dazugehörige Semantik, also eine Bedeutung, spricht man von Wissen. Information ist schliesslich Wissen in einer konkreten Situation, welches zur Lösung eines Problems benötigt wird. Datenbanken und Datenretrieval bewegen sich auf der Stufe Wissen, da die Daten mit einer Semantik (repräsentiert durch die Attribute eines Datensatzes) versehen sind. Information Retrieval Systeme liefern im Idealfall hingegen Informationen zur Lösung einer konkreten Problemstellung.

2.2. Wie funktioniert ein Information Retrieval System

Grundsätzlich ist ein Information Retrieval System (IRS) aus drei Teilen aufgebaut: Eingabe, Prozessor und Ausgabe. Als optionales viertes Element tritt ein Feedback seitens der Benutzer eines IRS auf. Rijsbergen³ veranschaulicht diesen Zusammenhang in der Abbildung 2:

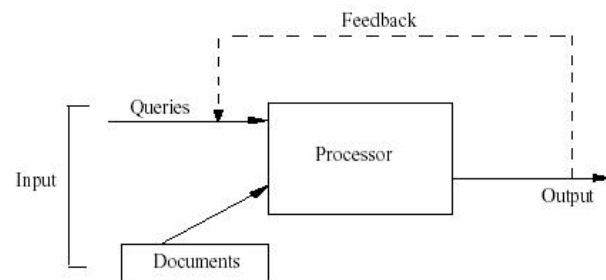


Abbildung 2 – Grundsätzlicher Aufbau eines IRS nach Rijsbergen. Als Eingabe (Input) für das IRS dienen einerseits Dokumente (Documents) und andererseits Abfragen (Queries). Diese beiden Elemente werden im Prozessor (Processor) – sozusagen dem Kern des IRS – verarbeitet und als Ausgabeergebnis (Output) dem Benutzer präsentiert. Der Benutzer reagiert darauf und verändert je nach der Qualität des Ergebnisses seine Anfrage ans IRS (Feedback).

2.2.1. Indexierung

Um Abfragen schnell bearbeiten zu können, verwenden die meisten Information Retrieval Systeme sogenannte Indizes. Diese beinhalten nicht den gesamten Inhalt der Dokumente, sondern speziell aufbereiteten Inhalt. Die Erstellung eines Indexes erfolgt grundsätzlich folgendermassen:

³ [VRIJS]

Als erstes werden die einzelnen Wörter aus einem Dokument identifiziert. Dies kann beispielsweise durch Leerzeichen und Interpunktionen geschehen. Anhand einer sogenannten Stopwortliste werden nun häufig auftretende Wörter, die nur geringe Bedeutung haben aussortiert. Es handelt sich dabei beispielsweise um Füllwörter, sprachliche Konstruktoren (wie "und") oder Präpositionen.

Luhn⁴ konnte nun zeigen, dass mittels des Zusammenhangs zwischen Rangreihenfolge von Wörtern in einem Text und der Auftretenshäufigkeit von Wörtern die Signifikanz von Wörtern und Sätzen ermittelt werden kann. Er veranschaulichte diese Methode anhand von Abbildung 3. Durch empirische Untersuchungen anhand von Dokumenten setzte er eine obere und untere Grenze fest ("Upper cut-off" respektive "Lower cut-off"), da sowohl zu häufig vorkommende Wörter wie auch solche mit sehr geringer Auftretenswahrscheinlichkeit nur eine geringe Signifikanz aufweisen. Diese Grenzen sind allerdings manuell festzulegen und lassen einen gewissen Spielraum für Ungenauigkeiten.

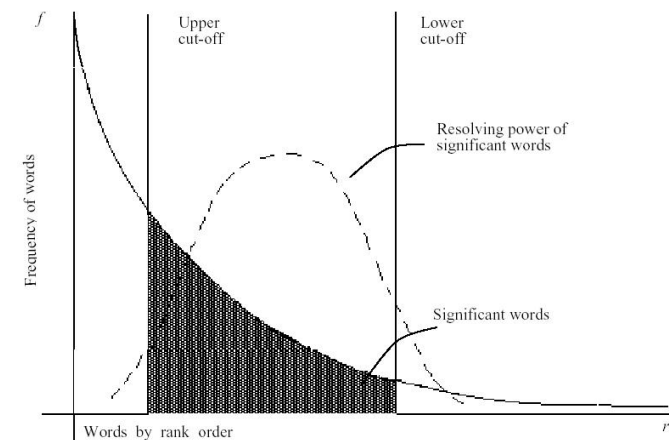


Abbildung 3 – Signifikante Wörter feststellen nach Luhn. Die X-Achse bezeichnet die Ranreihenfolge von Wörtern im Text, die Y-Achse die Auftretenswahrscheinlichkeit der Wörter. Die dunkel schraffierte Fläche ergibt die Menge der signifikanten Wörter.

Mittels eines weiteren Ansatzes lässt sich die Qualität der signifikanten Wörter noch erhöhen: Die Methode nennt sich Lemmatisierung: Dabei wird versucht, die Wörter auf ihre

⁴ [SCHLU]

Grundform (beispielsweise ein Verb auf den dazu gehörenden Infinitiv) oder den Wortstamm zurück zu führen. Je nach Sprache, in der die Dokumente vorliegen, ist dieses Vorgehen mehr oder weniger erfolgreich.

Um die einzelnen Wörter in ein Gebiet einordnen zu können, werden Thesauri eingesetzt. Die Enzyklopädie Encarta⁵ definiert den Begriff Thesaurus folgendermassen:

Thesaurus (griechisch thesauros: Schatz bzw. Schatzhaus), alphabetisch oder thematisch geordnete Zusammenstellung des Wortschatzes einer Sprache oder eines Themenbereichs.

[...] versteht man unter dem Begriff Thesaurus eine meist nach Themen geordnete Zusammenstellung aller sprachlichen und sonstigen Beziehungen eines Wortes innerhalb eines bestimmten Anwendungsbereiches (z. B. Medizin, Chemie, Wirtschaft etc.). Hier sind Thesauri folglich als Ordnungssysteme zu verstehen, die als grundlegende Hilfsmittel zur inhaltlichen Erschließung von Informationen über den jeweils gesuchten Begriff dienen.

Ein Thesaurus enthält normalerweise zu einem Term Über- und Unterbegriffe, Synonyme und Antonyme und themenverwandte Wörter.

Nachdem nun ein Dokument all diese Schritte durchlaufen hat und die Terme darin analysiert worden sind, kann es in einen Index eingefügt werden. Dieser ist zumeist mittels einer invertierten Liste realisiert. Eine invertierte Liste enthält eine Liste aller gefundenen Terme und zu jedem Term eine Relation zu den Dokumenten, in welchen der Term gefunden wurde. So können sehr schnell alle Dokumente gefunden werden, in denen ein gewisses Wort oder ein Term auftaucht. Die Abbildung 4 soll zeigen, wie ein solcher Index realisiert werden kann.

Die Arbeit des Indexierens kann entweder durch Menschen oder Roboter geschehen. Erstere Methode kommt bei Verzeichnissen wie Yahoo!⁶ zum Einsatz. Sie hat den Vorteil, dass die Qualität des Indexes sehr hoch ist. Dafür muss allerdings ein grosser Aufwand in Kauf

⁵ <http://encarta.msn.de/find/Concise.asp?z=1&pg=2&ti=721551729>

⁶ <http://www.yahoo.com>

genommen werden. Die meisten heutigen Suchmaschinen setzen deshalb automatische Crawler ein, die diese Arbeit übernehmen. Darauf wird speziell im Abschnitt 3.1.9 eingegangen.

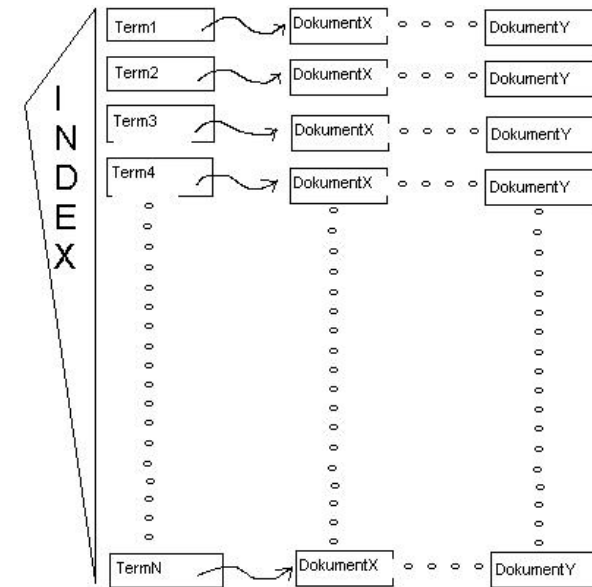


Abbildung 4 – Aufbau eines Indexes. Mittels einer invertierten Liste werden alle gefundenen Terme abgespeichert: Zu jedem Term wird auf die Liste der Dokumente verwiesen, in denen besagter Term vorkommt. Die invertierte Liste könnte beispielsweise auch so organisiert sein, dass sie eine Liste von Dokumenten enthält (Dokument1 bis DokumentN) und mit jedem Dokument verknüpft eine Liste der darin enthaltenen Terme.

2.2.2. Retrieval Modelle⁷

Es gibt sehr viele unterschiedliche Modelle, um Informationen mit einem Retrieval System wieder aufzufinden. Im folgenden wird auf die zwei grundlegenden Modelle, das Boolesche Modelle und das Vektorraummodell, eingegangen.

2.2.2.1. Boolesches Retrieval

Das Boolesche Retrieval ist historisch gesehen das älteste Modell für Information Retrieval. Es ist sehr einfach zu implementieren, liefert aber nicht sonderlich gute Ergebnisse. Die

⁷ [NFUHR]

Strategie liefert als Ergebnis die Dokumente, die für die Abfrage "wahr" sind. Es wird geprüft, ob der Term in der Abfrage in den Dokumenten enthalten ist oder nicht. Die Abfrage selbst wird mittels logischer Operatoren (beispielsweise AND, OR und NOT) gestaltet und verfeinert.

Vorteile:

- **Mächtigkeit:** Mittels der logischen Operatoren kann jede Teilmenge der Datenbasis gefunden werden.

Nachteile:

- **Antwortmenge:** Die Antwortmenge ist zu gross und schlecht strukturiert.
- **Gewichtung:** Die Dokumente werden nicht mit Gewichten versehen und können nicht in eine Reihenfolge der Relevanz gebracht werden.
- **Frageformulierung:** Die Formulierung der Abfrage gestaltet sich für viele Benutzer als kompliziert, da sie nicht an das Arbeiten mit logischen Operatoren gewöhnt sind.

Um den Nachteil der fehlenden Gewichtung zu überwinden, gibt es einen weiteren Ansatz, der auf dem Booleschen Retrieval basiert. Er nennt sich Fuzzy-Retrieval. Dabei wird zugelassen, dass die Dokumente in den Indexen mit Gewichten versehen werden. Dies ergibt im Gegensatz zum einfachen Booleschen Modell eine Rangordnung der Dokumente in der Antwort zu einer Abfrage. Doch auch wenn die Verwendung von Gewichtung einige Verbesserung in der Qualität der gelieferten Dokumente liefert, bleiben die anderen Probleme (Antwortmenge und komplizierte Frageformulierung) bestehen.

2.2.2.2. Das Vektorraummodell

Das Vektorraummodell wurde im Laufe des SMART-Projekts in Harvard und Cornell Anfang der 60er-Jahre von Gerard Salton entwickelt und in den 80er-Jahren von Wong und Raghavan überarbeitet.

Dabei werden die Frage und Dokumente als Punkte in einem mehrdimensionalen Vektorraum aufgefasst. Der Vektorraum wird durch die einzelnen Terme der Abfrage aufgespannt. Danach werden die Vektoren der gefundenen Dokumente auf Ähnlichkeit zum Vektor der Frage überprüft, was eine Reihenfolge der Dokumente ergibt. Die Abbildung von Bothe illustriert die Methode anhand einer Abfrage mit drei Termen und drei gefundenen

Dokumenten. Das Dokument₁ ist in diesem Beispiel der Frage (Vektor „Frage“) am ähnlichsten und erscheint im Abfrageergebnis an oberster Stelle.

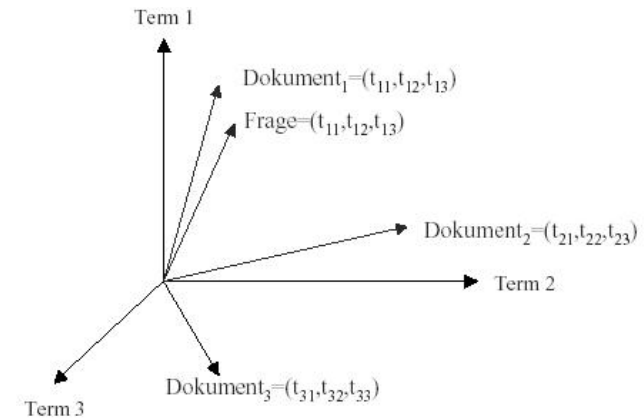


Abbildung 5 – Veranschaulichung des Vektorraummodells nach Bothe⁸. Jedes Dokument wird anhand des Auftretens der Terme 1 bis 3 durch einen Vektor symbolisiert. Diese werden mit dem Vektor der Frage verglichen und derjenige, der dem Frage-Vektor am nächsten liegt, wird im Abfrageergebnis zuerst ausgegeben.

Um die Retrievalqualität nochmals zu verbessern, wurden im Rahmen des SMART-Projektes heuristische Formeln zur Berechnung von Gewichten für die Dokumente bei der Indexierung entwickelt.

Vorteile:

- **Benutzerfreundlichkeit:** Das Modell ist anschaulich und erlaubt einfache Frageformulierungen.
- **Qualität:** Mit den oben erwähnten SMART-Gewichtungsformeln liefert das Vektorraummodell sehr gute Abfrageergebnisse.

Nachteile:

- **Heuristische Komponenten:** Die Gewichtung basiert lediglich auf heuristischen Komponenten. Es stellt sich die Frage, ob die Formeln auch noch gültig sind, wenn komplett andere Dokumente untersucht werden als während des SMART-Projekts.

⁸ [PBOTH]

- **Begründbarkeit des Modells:** Es ist theoretisch nicht zu begründen, warum zu einer Frage ähnliche Dokumente auch wirklich relevanten Inhalt haben sollen.

2.2.3. Messung⁹

Verschiedene Information Retrieval Systeme sollten verglichen werden können. Dafür werden klassischerweise drei Masse angewendet: Trefferquote, Vollständigkeit und Geschwindigkeit einer Abfrage. Dabei ist es zumeist so, dass sich die drei Masse so verhalten, dass sie sich entgegen wirken: Wird beispielsweise die Trefferquote eines IRS erhöht, leidet zumeist die Geschwindigkeit. Die Abbildung 6 verbildlicht diesen Zusammenhang.

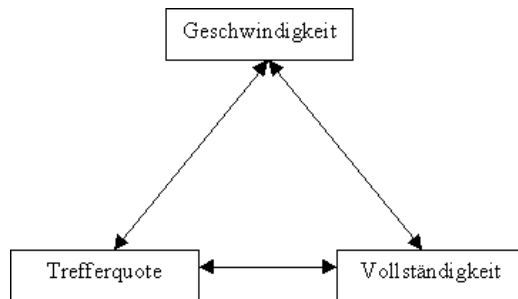


Abbildung 6 – Beziehung unter den drei Hauptmassen zur Beurteilung eines IRS. Sie wirken sich oft entgegen: Verändert man die Eigenschaften eines IRS bezüglich des einen Masse, hat dies (zumeist negative) Auswirkung auf die beiden anderen Masse.

2.2.3.1. Trefferquote (Precision)

Die Trefferquote ist ein Mass für die Güte der gelieferten Dokumente. Sie ist als das Verhältnis von relevanten Dokumenten zu den gefundenen Dokumenten definiert. Sie gibt also an, wie viele relevante Dokumente sich in allen gefundenen Dokumenten befinden. Erstrebenswert ist ein Wert möglichst nahe bei 1.

$$\text{Trefferquote} = \frac{\text{Anzahl relevante Dokumente}}{\text{Anzahl gefundene Dokumente}}$$

⁹ [KOKOI]

2.2.3.2. Vollständigkeit (Recall)

Die Vollständigkeit wird durch das Verhältnis von den gefundenen, relevanten Dokumenten zu der Anzahl aller relevanten Dokumente gemessen. Sie sagt demnach aus, wie viele relevante Dokumente gefunden wurden. Auch hier ist ein Wert nahe bei 1 erstrebenswert.

$$\text{Vollständigkeit} = \frac{\text{Anzahl gefundene, relevante Dokumente}}{\text{Anzahl aller relevanten Dokumente}}$$

2.2.3.3. Geschwindigkeit (Speed)

Was nützt einem Benutzer ein hochpräzises Information Retrieval System, wenn er sehr lange auf ein Abfrageergebnis warten muss. Deshalb ist das Geschwindigkeits-Mass eminent wichtig und ein entscheidender Faktor für die Akzeptanz und den Erfolg des Systems. Geschwindigkeit wird als die Zeit vom Bestätigen der Abfrage bis zur Ausgabe des Anfrageergebnisses durch das System definiert.

3. Information Retrieval im Web

Information Retrieval Systeme sind in verschiedensten Gebieten im Einsatz. Es ist klar, dass durch die Unterschiede in den Einsatzgebieten für das Information Retrieval im Web dessen Eigenheiten beachtet werden müssen: Denn zwischen dem Durchsuchen einer Bibliotheksdatenbank mittels eines IRS und dem Betreiben einer Suchmaschine für das Web liegen Welten. In den folgenden Abschnitten wird deshalb auf die Spezialitäten des Web eingegangen.

3.1. Spezialitäten des Web¹⁰

3.1.1. Datenmenge

Das World Wide Web umfasst eine riesige Menge an Daten. Der zahlenmässig grösste Teil davon sind Dokumente mit Text als Inhalt. Es kommen aber auch Multimedia-Inhalte wie Bilder, Audio-Dateien, Flash-Animationen, Videos und vieles mehr vor. Diese immense Datenmenge zu durchforsten und einigermaßen relevante Abfrageergebnisse zu generieren, stellt Information Retrieval Systeme fürs Web vor kaum zu lösende Aufgaben und Schwierigkeiten. Trotz einiger vielversprechender Ansätze ist der Anteil des "Dunklen

¹⁰ [LHUAN]

Webs", also des Bereiches des Web, der nicht mehr aufgefunden werden kann, sehr gross und tendenziell wachsend.

Die Abbildung 7 und Abbildung 8 sollen einen Überblick über das rasante Wachstum des Internet geben und andeuten, welche Datenmengen sich im Web verstecken. Dabei sind folgende Definitionen zu beachten:

- **Host:** Computer mit einer registrierten IP-Adresse
- **Web-Site:** Ein Web-Server. Ein Host kann mehrere Web-Server enthalten. Ebenso enthalten Web-Sites normalerweise eine grosse Anzahl an einzelnen Web-Seiten. Die faktisch enthaltenen Datenmengen einer Web-Site können demnach nur geschätzt werden.

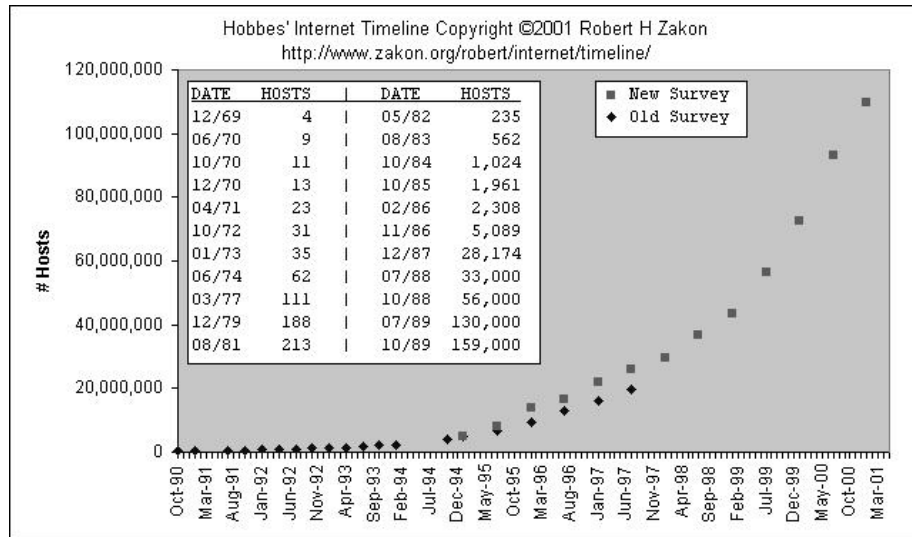


Abbildung 7 – Wachstum der Hosts nach Zakon

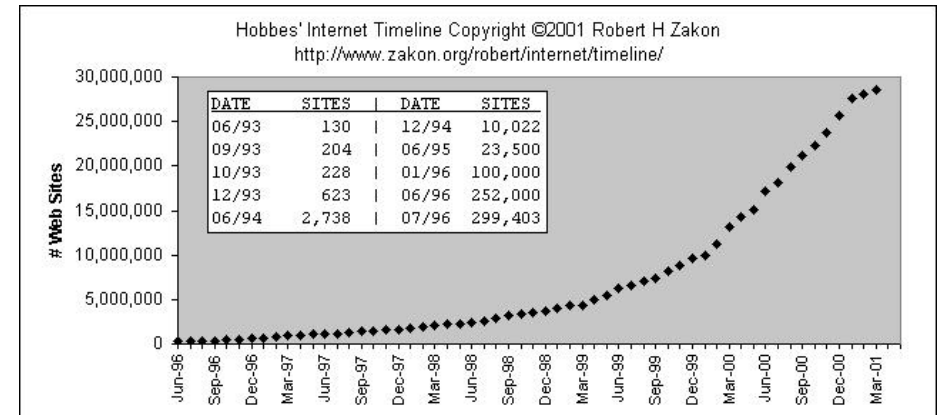


Abbildung 8 – Wachstum der Web-Sites nach Zakon

3.1.2. Dynamik

Indizes sind an und für sich für statische Datensammlungen optimiert. Nun verändert sich der Inhalt des Webs aber sehr schnell. Mit einer Aktualisierung des Indexes einmal pro Monat stösst der Benutzer immer wieder auf sogenannte tote Links. Das sind Dokumente, die entweder nicht mehr existieren oder deren Verknüpfung sich geändert hat. Vor allem bei Seiten, die einen häufigen Aktualisierungsrhythmus haben, ist ein monatliches Update des Indexes viel zu wenig.

3.1.3. Heterogenität der Daten

Heutige Web-Seiten bestehen nicht mehr nur noch aus Text, sondern enthalten grafische und animierte Elemente. Auch werden Audio-Inhalte angeboten. Oftmals werden die Seiten nicht statisch abgespeichert, sondern je nach Benutzereingabe dynamisch kreiert, indem auf eine Datenbank im Hintergrund zugegriffen wird. Traditionelle IRS sind dabei auf die Suche für einen speziellen Datentyp optimiert. Ein IRS fürs Web muss dagegen mit einer Fülle an verschiedensten Dateiformaten zurecht kommen und die verschiedenen Inhalte durchsuchen können.

Dieser Entwicklung entsprechend ist das Gebiet des Multimedia Retrieval ein Forschungsgebiet, dem grosse Beachtung geschenkt wird. Zum jetzigen Zeitpunkt muss aber gesagt werden, dass sich die im Betrieb befindlichen IRS hauptsächlich auf den textuellen Inhalt eines Dokumentes beschränken müssen.

Die Heterogenität der Daten lässt sich auch unter dem Gesichtspunkt der verschiedenen Gebiete betrachten: Im Web sind Dokumente zu unterschiedlichsten (Fach-)Gebieten auffindbar. Ein Begriff Mathematik hat beispielsweise im Web eine ganz andere Bedeutung und Relevanz als in einem anderen, spezialisierten Information Retrieval System. Ein solches kann sich gar nur mit dem Gebiet Mathematik befassen, womit natürlich eine Suchabfrage mit dem Wort Mathematik qualitativ ganz andere Ergebnisse liefert als dieselbe Anfrage, die an eine Suchmaschine für das Web gestellt wird.

3.1.4. Verschiedene Sprachen

Das Internet ist ein weltumspannendes Netz und theoretisch können alle Völker daran teilnehmen. Infolge des sogenannten Digital Divide (faktisch ist das Internet heute nur in den Industrieländern breitflächig verfügbar) hat aber der grösste Teil der Weltbevölkerung noch keinen Zugang zum Web. Trotz dieses Missstandes sind die auf dem Web erhältlichen Dokumente in verschiedensten Sprachen (mehr als 100) verfasst. Dies stellt IRS vor grosse Probleme: Wie soll ein Dokument, das beispielsweise in französisch abgefasst ist, von einer Suchmaschine gefunden werden, die eine Anfrage in deutschen Termen bearbeitet?

3.1.5. Heterogenität der Benutzer

Jeder Benutzer unterscheidet sich in seinen Erfahrungen, seinem Wissensstand, seinen Erwartungen – um nur einige wenige Punkte aufzuführen -, wenn er ein Information Retrieval System fürs Web bedient. Natürlich haben auch IRS für andere Einsatzgebiete mit diesem Problem zu kämpfen, aber da theoretisch die ganze Weltbevölkerung als potenzielle Benutzer eines IRS fürs Web angesehen werden muss, lässt sich der Grad der Benutzer-Heterogenität eines IRS fürs Web nicht mehr übertreffen! Ein einfaches Beispiel: Während ein Informatik-Student unter dem Wort "Java" in einem IRS fürs Web nach Artikeln zur Programmierung sucht, sucht eine andere Person vielleicht mit dem genau gleichen Begriff nach Reiseangeboten für die gleichnamige Insel.

3.1.6. Duplikate

Digitale Informationen lassen sich besonders einfach kopieren. Schätzungen besagen, dass es sich bei etwa 30% des verfügbaren Inhaltes auf dem Web um Duplikate handelt. Als Beispiel sei hier eine Software-Firma genannt, die eine Evaluationsversion eines Programms auf mehreren Servern (sogenannte Mirror-Sites) weltweit abspeichert. Ein Information Retrieval System steht nun vor der schwierigen Aufgabe, die doppelt vorhandenen Inhalte zu erkennen und nur einmal zu indexieren.

3.1.7. Hohe Verlinkung

Durchschnittlich enthält jede einzelne Webseite acht Links zu anderen Seiten. Für einen automatischen Indexer bringt dieser Umstand einerseits Vorteile mit sich: So können überhaupt andere Dokumente gefunden und indexiert werden. Aber er wird auch mit grossen Problemen konfrontiert: Ein Problem sind Zyklen in den Links, bei welchen der Indexer wieder an den Ausgangspunkt seiner Suche zurückkehrt. Ein anderes, viel allgemeineres Problem ist der damit verbundene Aufwand: Jede indexierte Seite impliziert durchschnittlich acht neue Indexierungsschritte.

3.1.8. Länge der Abfrageergebnisse und Benutzerverhalten

Bedingt durch die riesige Datenmenge, die teilweise recht simplen Retrieval Modelle und offen formulierten Abfragen ist die Chance gross, dass sich ein Benutzer einer Unzahl zurückgelieferter Dokumente gegenüber sieht. Wünschenswert wäre hingegen eine kleine Menge an relevanten Dokumenten. Niemand ist bereit, 200 Dokumente (was noch nicht mal eine grosse Anzahl darstellt, wenn man die Zahl auf die Gesamtheit der indexierten Dokumente betrachtet) durchzusehen. Verschiedene Studien haben denn auch gezeigt, dass 85% der Benutzer von Suchmaschinen nur die erste Seite mit Abfrageergebnissen durchsehen.

Information Retrieval Systeme fürs Web müssen also besonders viel Wert auf das Benutzerverhalten legen. Dies zeigen auch verschiedene Studien¹¹ zu diesem Punkt. Als Beispiele hierfür seien diejenige von NPD New Media Service zum Thema Such- und Portalsite und die Studie des SIGIR Forums zu den Benutzeranfragen auf dem Web.

¹¹ [SENGI]

eventuell mit logischen Operatoren wie AND, OR etc. versehen – enthalten oder natürlichsprachlich formuliert sein. Heute ist die zweite Methode allerdings noch recht selten. Als Beispiele für Suchmaschinen können Google¹⁵ oder AltaVista¹⁶ genannt werden.

3.2.2. Metasuchmaschinen

Metasuchmaschinen sind zumeist nicht selbst mit der ganzen Architektur und Funktionalität einer eigenständigen Suchmaschine ausgestattet. Sie bieten dem Benutzer aber gleichermassen die Möglichkeit, eine selbst formulierte Anfrage einzugeben. Diese wird nun an mehrere Suchmaschinen weitergeleitet. Der grosse Vorteil einer Metasuchmaschine ist, dass sie aus den Ergebnissen der verschiedenen Suchmaschinen mittels spezialisierten Algorithmen besser in der Lage ist, relevante Dokumente rauszufiltern und dem Benutzer damit eine höhere Retrieval-Qualität zu bieten. Als Beispiel für Metasuchmaschinen sei hier MetaCrawler¹⁷ angegeben.

3.2.3. Agenten

Agenten sind im Grundsatz Programme, die autonom einer Aufgabe nachgehen und damit einen Auftrag bearbeiten. Da sie mehrheitlich selbständig arbeiten, fliessen bei der Entwicklung von Agenten auch Ansätze aus dem Gebiet der künstlichen Intelligenz ein. Der Agent ist für ein bestimmtes Einsatzfeld konstruiert und kann sich in diesem mehr oder weniger intelligent bewegen und auf seine Umwelt reagieren. Als Beispiel für einen Agenten kann ResearchIndex¹⁸ angegeben werden: Dieser ist spezialisiert auf wissenschaftliche Publikation und automatisiert viele Arbeiten wie beispielsweise die Textanalyse von PDF-Dokumenten auf Zitate.

3.2.4. Kataloge

Verzeichnisse oder Kataloge sind weit verbreitet als Information Retrieval Systeme für das Web. Wichtigster Unterschied zu den Suchmaschinen ist, dass hier die Dokumente nicht automatisch durch Roboter indexiert werden, sondern zuerst von Redaktoren geprüft werden und in einen Katalog eingefügt werden. Diese erlauben es einem Benutzer, schnell über einen allgemeinen Themenbereich und darin durch weitere Verfeinerung schnell Informationen mit hoher Relevanz zu finden. Kataloge eignen sich vor allem für standardisierte

¹⁵ <http://www.google.com>

¹⁶ <http://www.altavista.com>

¹⁷ <http://www.metacrawler.com>

¹⁸ <http://www.researchindex.com>

Informationsangebote. Sie müssen bedingt durch die Dynamik des Webs in mühsamer Einzelarbeit gepflegt werden. Zudem erfordert die Arbeit des Indexierens ein ganzes Heer an Redaktoren, um ein einigermaßen vollständiges Verzeichnis zu erhalten. Der berühmteste Vertreter dieser Art von IRS ist sicherlich Yahoo!¹⁹.

4. Beurteilung

Information Retrieval im Web unterscheidet sich in weiten Bereichen stark vom Datenretrieval. Die Unterschiede sind vor allem auf die Eigenheiten des Web, aber auch auf das veränderte Benutzerverhalten und das sehr breit gefächerte Zielpublikum zurück zu führen.

Deshalb ist es sehr wichtig, dass sich IRS für das Web an den speziellen Rahmenbedingungen orientieren und auf den Benutzer mit seinen Gewohnheiten eingehen. Es gibt heute viele Ansätze, die für die bestehenden Probleme des Information Retrieval im Web gute Lösungen bieten.

Ob Verzeichnisse, (Meta-)Suchmaschinen oder Agenten, alle haben ihre Berechtigung: Je nach Art der gesuchten Information kann die eine oder andere Methode bessere Ergebnisse liefern. Optimal ist natürlich eine Kombination mehrerer Vorgehensweisen. Es darf aber nicht unerwähnt bleiben, dass Information Retrieval im Web ein Gebiet ist, dass sich ständig weiter entwickelt – gerade weil noch viele Schwierigkeiten ungelöst sind.

¹⁹ <http://www.yahoo.com>

5. Literaturverzeichnis

Schriftliche Quellen:

- [VRIJS]: Van Rijsbergen, C. J. (1979): Information Retrieval. London.
- [SCHLU]: Schultz, C.K. / Luhn, H.P. (1968): Pioneer of Information Science. London.
- [LHUAN]: Huang, L. (2000): A Survey On Web Information Retrieval Technologies. New York.
- [KOKOI]: Kobayashi, M. / Takeda, K. (2000): Information Retrieval on the Web. Tokyo.
- [BRIPA]: Brin, S. / Page, L. (1998): The Anatomy of a Large-Scale Hypertextual Web Search Engine. Stanford.
- [MLAGE]: Lager, M. (1996): Spinning a Web Search. Santa Barbara.
- [JSBAS]: Jansen, B. J. / Spink, A. / Bateman, J. / Saracevic, T. (1998): Real life information retrieval: A study of user queries on the Web. Melbourne.
- [NFUHR]: Fuhr, N. (2000): Information Retrieval – Skriptum zur Vorlesung im WS 00/01. Dortmund.
- [PSCHA]: Schäuble, P. (1992): A Tutorial on Information Retrieval. Zürich.
- [BARIB]: Baeza-Yates, R. / Ribeiro-Neto, B. (1999): Modern Information Retrieval. New York.

Elektronische Quellen (Juni 2001):

- [AWICH]: Wichmann, A. (1999): Aufbau und Techniken von Suchmaschinen für das WWW.
<http://www-student.informatik.uni-bonn.de/~wichmann/writings/webcrawlers/index.html>
- [PBOOTH]: Bothe, P. (2000): Text Retrieval.
http://www.in.tu-clausthal.de/~hoerner/hs_datenbanken/VortragBothe.pdf
- [BROHE]: Broder, A. / Henzinger M. (1998): Information retrieval on the Web.
<http://www.research.compaq.com/SRC/personal/broder/focs98/ppframe.htm>
- [GKNOR]: Knorz, G. (1994): Automatische Indexierung.
<http://www.iud.fh-darmstadt.de/iud/wwwmeth/publ/skript/autind94/paper1.htm>
- [BSCHU]: Schulzki, B. (2000): Textbasiertes Ranking.
<http://www.informatik.hu-berlin.de/~schulzki/sm/sm1.html>
- [SENGI]: Search Engine Reviews, Ratings & Tests.
<http://www.searchenginewatch.com/reports/index.html>