



**Universität Karlsruhe (TH)**  
Fakultät für Informatik  
*Institut für Programmstrukturen und  
Datenorganisation (IPD)*

**Hauptseminar**  
**Spezifikations- und Selektionsmethoden für Daten und  
Dienste - S2D2**

# **Verfahren des Information Retrieval**

**Seminararbeit**

von

**Maxim Jochim**

Wintersemester 2006/2006



# Inhaltsverzeichnis

<b>1</b>	<b>Grundlegendes</b>	<b>4</b>
1.1	Information Retrieval . . . . .	4
1.2	Beispiel einer Datenbankrecherche . . . . .	5
1.3	Daten - Information - Wissen . . . . .	6
<b>2</b>	<b>Verfahren</b>	<b>7</b>
2.1	Bool'sches Retrieval . . . . .	7
2.2	Fuzzy-Retrieval . . . . .	8
2.3	Das Vektorraummodell . . . . .	9
<b>3</b>	<b>Gewichtungsmethoden</b>	<b>11</b>
3.1	Häufigkeitsformeln . . . . .	12
3.2	Inverted Document Frequency . . . . .	13
<b>4</b>	<b>Schlusswort</b>	<b>13</b>



## Zusammenfassung

Mit der zunehmenden Verbreitung des Internets vergrössert sich auch die Menge der Information, die uns prinzipiell zur Verfügung steht. Allerdings wird es auch zunehmend schwieriger an bestimmte, mit der momentanen Situation im Zusammenhang stehende, Information heran zu kommen. Daher ist man ständig bemüht aus dieser Informationsflut gezielt benötigte Information heraus zu Filtern.

Im Rahmen dieser Studienarbeit werden hier einige nicht probabilistische Modelle des Information Retrieval vorgestellt, die eine sehr grosse Verbreitung in den Information Retrieval Systemen gefunden haben. Als erstes müssen aber die Daten von der Information unterscheiden werden um einen klaren Informationsbedarf, in Form einer Anfrage an das Information Retrieval System, definieren zu können. Da es in manchen Fällen aber nicht klar ist, welches der Ergebnisdokumente mehr und welche weniger relevante Information enthält, wird es hier auch auf die Problematik der Informationsrelevanz des gelieferten Ergebnisses eingegangen.

# 1 Grundlegendes

## 1.1 Information Retrieval

„Inhaltliche Suche in Texten“ wäre der passendste Ausdruck, wenn man Information Retrieval (IR) mit wenigen Worten beschreiben will. In Wirklichkeit wird damit aber nur *ein* wesentlicher, aber auch der wichtigste, Bereich des Information Retrieval beschrieben, den man auch als Textretrieval oder Dokumentenretrieval bezeichnet.

Das klassische Anwendungsgebiet des IR sind die Recherchen in den Literaturdatenbanken, die heutzutage in Form der Digitalen Bibliotheken zunehmend an Bedeutung gewinnen. IR ist besonders populär geworden durch die Anwendung der Suchmaschinen im Internet. Dadurch diese kommt jeder Internet-Nutzer mit IR-Methoden in Berührung. Neben der Suche in Texten werden auch mehr und mehr IR-Anwendungen für multimediale Daten realisiert, wobei insbesondere Methoden des Bildretrieval, am Beispiel des [images.google.de](http://images.google.de), eine zunehmende Verbreitung finden.

Jeder, der eine dieser Anwendungen wiederholt genutzt hat, merkt die wesentlichen Unterschiede zwischen IR-Anwendungen und den klassischen Datenbanksystemen:

- In IR-Systemen bereitet die Formulierung einer zum aktuellen Informationsbedürfnis passenden Anfrage erhebliche Probleme.

- Meistens durchläuft der Prozess der Anfrageformulierung mehrere Iterationen, bis passende Antworten gefunden werden.
- Anfragen liefern potentiell sehr viele Antworten (vgl. die Gesamtzahl der Treffer bei Internet-Suchmaschinen), aber nur wenige davon sind für den Nutzer interessant.
- Das vorgenannte Problem entschärft sich durch die vom System bereitgestellte Rangordnung der Antworten, wodurch potentiell relevante Antworten gehäuft am Anfang der Rangliste auftauchen (z.B. betrachten bei Internet-Suchmaschinen mehr als 90% aller Nutzer nur die ersten 10 Antworten)
- Bei Textdokumenten, aber noch stärker bei Bildern zeigt sich, dass die systemintern verwendete Repräsentation des Inhalts von Dokumenten teilweise in keinem Zusammenhang stehen, auf jeden Fall aber mit Unsicherheit behaftet ist.

## 1.2 Beispiel einer Datenbankrecherche

Angenommen, wir suchen Literatur zum Stand der Forschung im Bereich der alternativen Energien für den Einsatz in den Personenkraftwagen, wobei uns insbesondere interessiert, was man bei BMW und nicht bei Mercedes erreicht hat.

Die Beispieldatenbank AUTOS besteht aus Dokumenten, die Artikel beschreiben, indem sie neben bibliographischen Angaben eine Kurzzusammenfassung (Abstract), eine Einordnung in ein hierarchisches Indexsystem und Stichwörter enthalten. Auf die Eingabe von (Kombinationen von) Wörtern liefert das (Bool'sche) Retrievalsystem von AUTOS die Dokumente, die die Wörter in der angegebenen Form enthalten.

Deshalb überlegen wir uns zunächst, welche Stichwörter das Problem besonders gut beschreiben. Das sollten Wörter sein, die spezifisch für die Fragestellung sind, aber dann doch wieder so allgemein, dass man annehmen kann, dass sie in jedem Artikel, der für uns wichtig ist, vorkommen.

Eine Anfrage mit Stichwörtern „Personenkraftwagen“, „alternative Energie“, „BMW“ und „Mercedes“ könnte zum Beispiel so aussehen: *PERSONENKRAFTWAGEN and ALTERNATIVE ENERGIE and BMW and not MERCEDES*. Sie wird vom Retrieval System so interpretiert: Suche alle Dokumente, in denen die Zeichenketten „PERSONENKRAFTWAGEN“, „ALTERNATIVE ENERGIE“ und „BMW“, aber nicht „MERCEDES“ irgendwo im Text vorkommen. Dabei wird zwischen Gross- und Kleinbuchstaben nicht

unterschieden, und das Leerzeichen zwischen ALTERNATIVE und ENERGIE kann auch ein anderer „white space“ sein (also z. B. ein Zeilenumbruch, auch in Verbindung mit mehreren Leerzeichen)

### 1.3 Daten - Information - Wissen

Datenbanksysteme enthalten Daten. IR-Systeme sollen die Suche nach Information<sup>1</sup> unterstützen. Daher stellt sich die Frage, ob IR-Systeme Information enthalten? Schließlich ist vor allem in KI-Bereich (Künstliche Intelligenz) häufig die Rede von Wissensbasen. Was ist denn nun der Unterschied zwischen Daten, Wissen und Information?

Die Daten an sich sind auf der syntaktischen Ebene angesiedelt. In diesem Sinne wäre also eine Datenbasis eine nackte Sammlung von Werten ohne jegliche Semantik. Kommt eine verwertbare und interpretierbare Semantik hinzu, so sprechen wir von Information [Wik]. Dementsprechend enthalten also Datenbanksysteme nicht nur Daten, sondern auch Information, weil zusätzlich zu den Daten zumindest ein Teil der Semantik des jeweiligen Anwendungsgebietes auch im System modelliert wird. Genauso enthält jedes IR-System Information (im Gegensatz etwa zu dem Fall, wo man Texte einfach in einer Datei abspeichert und mit Hilfe eines Texteditors durchsucht).

Wissen schließlich ist auf der pragmatischen Ebene definiert und könnte so formuliert werden: „Die Gesamtheit aller organisierten Informationen und ihrer wechselseitigen Zusammenhänge, auf deren Grundlage ein vernunftbegabtes System handeln kann“ [Wik]. Das Wissen ist also die Information, die von jemandem in einer konkreten Situation zur Lösung von Problemen benötigt wird. Da dieses Wissen häufig nicht vorhanden ist, wird danach in externen Quellen gesucht. Hierbei dient ein Informationssystem dazu, aus der gespeicherten Information das benötigte Wissen zu extrahieren. Wir sprechen auch von Informationsflut, wenn uns grosse Mengen an Information zugeleitet werden, aus denen wir nur mit Mühe das benötigte Wissen extrahieren können. Daher sind wir auch oft bereit, für gezielt bereitgestelltes Wissen zu zahlen (z.B. Tageszeitung, werbefreies Fernsehen). Schlagwortartig lässt sich die Beziehung zwischen Information und Wissen durch folgende Formulierung ausdrücken „Wissen ist Information in Aktion“. Um dieses Wissen zu extrahieren bedient man sich der IR-Systeme, die die Information durch verschiedene Retrieval-Verfahren, die Gegenstand dieser Arbeit sind, aufbereiten und das benötigte Wissen in der konkreten Situation bereitstellen.

---

<sup>1</sup>Da Information schlecht Quantifizierbar ist bleibt sie auch im Plural „Information“

## 2 Verfahren

Der Rahmen dieser Arbeit erlaubt es mir nicht auf alle Verfahren des Information Retrieval einzugehen, deswegen werde ich hier lediglich die grundlegenden Verfahren erläutern. Auf diesen basieren weitere Verfahren die mit dem Ziel der Verbesserung der Retrievalqualität entwickelt wurden. Anfangen möchte ich mit dem einfachsten, mengenbasierten, logischen Verfahren der auf der Bool'schen Algebra aufbaut.

### 2.1 Bool'sches Retrieval

Bool'sches Retrieval ist historisch als erstes Retrievalmodell entwickelt und eingesetzt worden. Es zeichnet sich durch logische Klarheit aus doch liefert auf eine Anfrage eine ungeordnete Menge von Dokumenten. Eingesetzt wurde es anfangs um Retrieval mit Hilfe von Schlitzlochkarten durchzuführen. Auch als man später die Dokumente auf Magnetbändern speicherte, war bool'sches Retrieval das einzig anwendbare Modell, denn aufgrund der geringen Speicherkapazität damaliger Rechner musste direkt nach dem Einlesen des Dokumentes entschieden werden, ob es als Antwort ausgedruckt werden sollte oder nicht. Obwohl sich die Rechnerhardware seitdem rasant weiterentwickelt hat, hat man in der Praxis dieses Modell bis heute nicht grundlegend in Frage gestellt.

Grundidee des Bool'schen Retrievals sind einfache Mengenoperationen auf Mengen von Dokumenten, die durch Attributwerte der Dokumente charakterisiert sind (z.B. das Auftreten der Terme im Dokument). Eine Anfrage ist eine Verknüpfung von Paaren aus einem Attribut und einem dazugehörigen Attributwert (Attribut-Wert-Paare). Ein Attribut-Wert-Paar steht in der Anfrage für die Menge der Dokumente, bei denen das entsprechende Attribut den angegebenen Wert annimmt. Ein Attribut ist dabei eine Abbildung, die einem Dokument einen Wert zuordnet. Bezeichne  $D$  die Menge aller Dokumente,  $T = \{t_i \mid i = 1, \dots, n\}$  die Menge der Indexterme und  $t : D \rightarrow T$ ,  $t(d) = t_i$  ein Attribut, so ist

$$D_{t,t_i} = t^{-1}(t_i) = \{d \in D \mid t(d) = t_i\}$$

die Menge der Dokumente, die durch Attributwert  $t_i$  charakterisiert sind. Wenn in einer Anfrage zwei Attribut-Wert-Paare  $(t, t_i)$   $(s, s_i)$  durch bool'sche Operatoren verknüpft sind, so bezeichnen sie eine entsprechende Verknüpfung der Dokumente, z. B.  $(t, t_i)$  AND  $(s, s_i)$  liefert den Durchschnitt  $D_{t,t_i} \cap D_{s,s_i}$ .

Speziell für Textdokumente, z. B. in einer Literaturdatenbank, sind die wichtigsten Attribute das Auftreten der Terme in den verschiedenen Feldern

der Dokumente. Ein Attribut  $TI_{t_1} : D \rightarrow \{false, true\}$  wäre z.B. das Auftreten des Terms  $t_1$  im Titelfeld des Dokumentes.

$$D_{TI_{t_1}, true} = \{d \in D \mid d \text{ enthält } t_1 \text{ im Titelfeld}\}$$

Wäre durch  $TI_{t_1}$  charakterisierte Dokumentmenge, die diejenigen Dokumente enthält, bei denen  $t_1$  im Titelfeld vorkommt. Definiert man  $x$  als die Menge der Terme, die im Titelfeld des Dokuments vorkommen, dann ergibt sich die durch Term  $t_1$  charakterisierte Dokumentmenge als

$$D_{x,t_1} = x^{-1}(\{t_1\}) = \{d \in D \mid d \text{ enthält } t_1 \text{ im Titelfeld}\}$$

## Beurteilung

Das bool'sche Modell wurde in vielen Veröffentlichungen (z.B. [Bel61]) für ziemlich ungeeignet für die Anwendung im IR befunden, es wurden unter anderem folgende Nachteile des bool'schen Retrieval genannt:

1. Die Grösse der Antwortmenge ist schwierig zu kontrollieren.
2. Keine Ordnung der Ergebnisse nach Relevanz.
3. Trennung in „gefunden“ und „nichtgefunden“ zu streng:  
Z. B. zu  $q = t_1 \wedge t_2 \wedge t_3$  werden Dokumente mit zwei gefundenen Termen genauso verworfen wie mit 0.
4. Anfrageformulierung ist sehr umständlich
5. Retrievalqualität ist schlechter als von anderen Modellen (Fuzzy-Retrieval, gewichtete Indexierung)

## 2.2 Fuzzy-Retrieval

Als ein Ansatz, um einige der Nachteile von bool'schem Retrieval zu überwinden, wurde das Fuzzy-Retrieval vorgeschlagen. Im Unterschied zum bool'schen Modell werden hier bei den Dokumenten-Beschreibungen auch gewichtete Indexierungen zugelassen, d.h.  $d_{t_i} \in [0, 1]$ . Frage-Beschreibungen und Retrievalfunktion sind wie beim bool'schen Retrieval definiert. Durch die gewichtete Indexierung liefert die Retrievalfunktion jetzt Werte  $\rho(q_k^d, \vec{d}_t) \in [0, 1]$ . Damit ergibt sich im Gegensatz zum bool'schen Modell nun eine Rangordnung der Antwortdokumente und die diesbezüglichen Nachteile des bool'schen Retrieval entfallen. Aber auch in diesem Fall erwies sich die Retrieval-Funktion als

ungünstig. Um dies zu zeigen konstruieren wir folgendes Beispiel:

$$\begin{aligned}T &= \{t_1, t_2\} \\q &= t_1 \wedge t_2 \\ \vec{d}_1 &= (0.4, 0.4) \quad , \quad \vec{d}_2 = (0.39, 0.99) \\ \rho(q, \vec{d}_1) &= 0.4 \quad , \quad \rho(q, \vec{d}_2) = 0.39\end{aligned}$$

Bezüglich  $t_2$  hat  $d_1$  einen niedrigeren Indexgewicht, doch wegen der Verwendung der Minimum-Funktion bei der Konjunktion ist das höhere Gewicht des  $t_1$  ausschlaggebend für das insgesamt höhere Retrievalgewicht von  $d_1$ .

### Beurteilung

Zusammengefasst bietet Fuzzy-Retrieval folgende Vor- und Nachteile

- + Rangordnung der Dokumente durch gewichtete Indexierung
- Retrievalqualität ist immer noch schlecht im Vergleich zu, VR-Modell
- Umständliche Frageformulierung wie beim bool'schen Retrieval

## 2.3 Das Vektorraummodell

Das Vektorraummodell (VRM) ist wahrscheinlich das bekannteste Modell im Information Retrieval. Es wurde ursprünglich im Rahmen der Arbeiten am SMART-Projekt entwickelt [Sal71].

Im VRM werden Dokumente und Fragen als Punkte in einem Vektorraum aufgefasst, der durch die Terme der Datenbasis aufgespannt wird. Beim Retrieval wird dann nach solchen Dokumenten gesucht, deren Vektoren ähnlich zum Fragevektor sind. Durch diese geometrische Interpretation ergibt sich ein sehr anschauliches Modell. Der zugrundeliegende Vektorraum wird als orthonormal angenommen (eine starke Vereinfachung zu den realen Verhältnissen), d.h.:

- alle Term-Vektoren sind orthogonal (und damit auch linear unabhängig), und
- alle Term-Vektoren sind normiert.

Die im VRM zugrundegelegte Dokument-Beschreibung ist ähnlich der des Fuzzy-Retrieval eine gewichtete Indexierung.

$$d_t^D = \vec{d}_t \text{ mit } d_{t_i} \in \mathbb{R} \text{ für } i = 1, \dots, n$$

Die Beschreibung der Frage hat die gleiche Struktur:

$$q_k^Q = \vec{q}_k \text{ mit } q_{k_i} \in \mathbb{R} \text{ für } i = 1, \dots, n$$

Als Retrieval Funktion werden dabei verschiedene Vektor-Ähnlichkeitsmaße, z. B. Cosinus-Maß, Overlap-Maß, Jaccard-Maß, angewendet doch meistens wird mit dem Skalarprodukt gearbeitet.

$$\rho(\vec{q}_k, \vec{d}_t) = \vec{q}_k \cdot \vec{d}_t$$

An folgendem Beispiel soll die Anwendung des Modells veranschaulicht werden: Fragestellung: „Alternative Energie in den Personenkraftwagen von BMW, nicht von Mercedes“

$t_i$	$q_{k_i}$	$d_{1_i}$	$d_{2_i}$	$d_{3_i}$	$d_{4_i}$
Alternative Energie	3	1	0,5		1
in den			1	1	0,5
Personenkraftwagen	1	1		1	
von			0,5	1	0,5
BMW	2	1		1	1
¬Mercedes	-1		1	1	
<b>Retrievalgewicht</b>		6	0,5	2	5

Die Dokumente werden in einer Reihenfolge, die den Retrievalgewichten entspricht ausgegeben, also  $d_1, d_4, d_3, d_2$ .

## Beurteilung

Zusammengefasst bietet das VRM folgende Vor- und Nachteile:

- + Das VRM ist ein relativ einfaches, anschauliches Modell, das wegen der einfachen Art der Frageformulierung auch benutzerfreundlich ist.
- + Das Modell ist unmittelbar auf neue Kollektionen anwendbar, im Gegensatz zu den probabilistischen Modellen, wo das Sammeln der Relevance-Feedback-Daten teilweise erforderlich ist
- + Das Modell liefert in Kombination mit Gewichtungsformeln eine sehr gute Retrievalqualität.
- Wegen des heuristischen Ansatzes bei der Indexierungsgewichtberechnung ist die Erweiterung der Dokumentrepräsentation nur bedingt durchführbar, da eventuell erst umfangreiche Experimente durchgeführt werden müssen um die geeignete Gewichtungsformel zu finden. (Z. B. stärkere Gewichtung der Terme im Titelfeld)

- In dem Modell wird kein Bezug auf die Retrievalqualität genommen, es ist also theoretisch nicht zu begründen, warum die zu einer Frage ähnlichen Dokumente auch relevant sein sollen.

### 3 Gewichtungsmethoden

Bis jetzt vorgestellten Ähnlichkeitsfunktionen für Dokument- und Anfragevektoren wurden betrachtet, ohne vorher zu sagen, woher die Gewichte, die in den Vektoren stehen, kommen. Nur beim Bool'schen Retrieval kamen die Werte *false, true* vor in Abhängigkeit davon ob ein Term in einem Dokument oder einer Anfrage vorkam oder nicht. Hier zeige ich einige Methoden, mit denen Gewichtsvektoren für Dokumente und Anfragen bestimmt werden können.

Die einfachste Methode wäre die Gewichtungen von Termen bei der Indexierung von Menschen eingegeben zu lassen was bei einer grossen Dokumentenmenge ziemlich lange dauern würde. Auch der Anfragende könnte die Terme bei der Anfrage gewichten, wobei diese Gewichtungen durch Feedbackmethoden verfeinert werden können. In diesen Fällen sind die Gewichtungen von dem jeweiligen Kontext abhängig, also bei der Indexierung von dem Objekt bzw. Dokument, das indexiert wird, bei der Anfrage von dem Informationsbedürfnis der Anfragenden.

Es gibt aber auch Kriterien für die Gewichtung von Termen, die nicht vom Kontext abhängig sind. So sind zum Beispiel allgemeine Terme wie „Ergebnis, Methode, Verfahren, Zusammenfassung, ...“ schlechte Suchterme, da sie nicht an ein bestimmtes Fachgebiet gebunden sind, während Terme, die nur in bestimmten Wissensgebieten vorkommen, gute Terme sein sollten. Es sei denn, man sucht explizit nach einem Verfahren, Methode oder Zusammenfassung, oder möchte sich erst einen Überblick verschaffen um später seine Suche zu verfeinern. In einer Näherung an dieses Problem kann die Häufigkeit von Termen in Dokumenten betrachtet werden. Terme, die in sehr vielen Dokumenten vorkommen, haben eine schlechtere Diskriminationskraft als Terme die in weniger Dokumenten vorkommen. Andererseits werden mit Termen, die nur in sehr wenigen Dokumenten vorkommen im allgemeinen keine umfassenden Suchergebnisse erzielt werden können. Diesen Zusammenhang veranschaulicht Abbildung 1.

*Häufige Terme können beim Retrieval durch eine Stoppwortliste ausgeschlossen oder durch eine schwache Gewichtung (z.B. IDF) abgeschwächt werden. Seltene Terme werden meistens nicht gesondert behandelt, d.h., die rechte Trennlinie wird in der Regel ignoriert. [Fer03]*

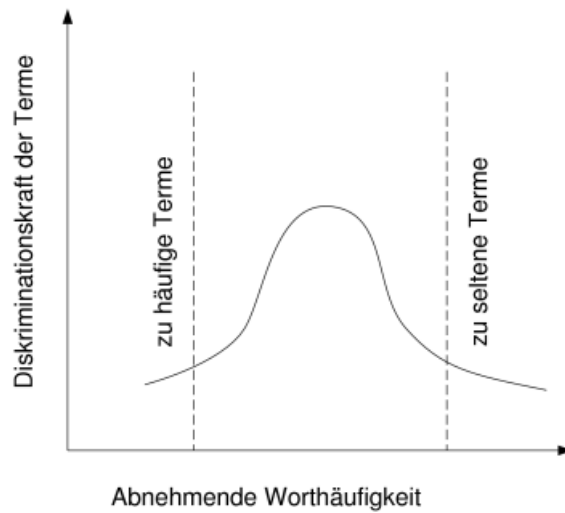


Abbildung 1: Diskriminationskraft von Termen.

Neben dieser allgemeinen Häufigkeit auch der Kontext des Dokument berücksichtigt werden. So kann die Wichtigkeit eines Terms im Dokument durch die Häufigkeit seines Vorkommens im Dokument ausgedrückt werden.

Die nachfolgenden Formeln wurden in den experimentellen Systemen verwendet. Für das bessere Verständnis dieser definieren wir  $h(i, j)$  als die Häufigkeit eines Terms  $t_j$  im Dokument  $d_i, d(j)$  die Anzahl der Dokumente in denen Term  $t_j$  vorkommt,  $D = (d_1, \dots, d_m)$  und  $T = (t_1, \dots, t_n)$  sind wieder die Mengen der Dokumente und Terme, wobei  $m$  und  $n$  ihre Größen ausdrücken.

### 3.1 Häufigkeitsformeln

Formeln, die die Häufigkeit eines Terms in einem Dokument berücksichtigen

$$h(i, j)$$

$$\frac{h(i, j)}{d(j)}$$

$$\frac{h(i, j)}{1 + h(i, j)}$$

### 3.2 Inverted Document Frequency

Formeln, die die Häufigkeit eines Terms in einer Dokumentmenge berücksichtigen (**IDF=Inverted Document Frequency**) Ein Term, das nur in wenigen Dokumenten oft vorkommt ist geeigneter als eines, dass in fast jedem Dokument oder nur sehr gering auftritt.

$$\ln \left( \frac{m}{d(i)} \right)$$
$$\ln \left( \frac{m - d(i)}{d(i)} \right)$$

Diese heuristische Formeln wurden bei den Arbeiten am experimentellen Retrieval System SMART [McG83] zur Berechnung der Indexierungsgewichte für Dokumente (und Fragen) entwickelt. Sie erwiesen sich als besonders leistungsfähig und wurden später im Rahmen der Arbeiten zu den experimentellen Systemen „Inquery“ (University of Massachusetts) und „OKAPI“ (MS Research Lab Cambridge) weiterentwickelt.

## 4 Schlusswort

Im Rahmen dieser Seminararbeit habe ich die grundlegenden Information Retrieval Verfahren vorgestellt und ihre Vor- und Nachteile erläutert. Ob Text-, Bildersuche oder Datenbankrecherche, die Einsatzgebiete der Information Retrieval sind vielfältig und die oben aufgeführten Verfahren werden heutzutage sehr breit eingesetzt. Mit der zunehmenden Internetverbreitung und dem Einsatz der Internetsuchmaschinen gewannen diese Verfahren stark an Bedeutung. Manche Verfahren werden in der Literatur als eigenständige präsentiert, doch bei der näheren Betrachtung erweisen sich diese oft als eine Abwandlung des einen oder des anderen bekannten Verfahrens. Abschließend bleibt noch zu erwähnen, dass Information Retrieval ein junges Forschungsgebiet ist auf dem noch viel, insbesondere zu Verbesserung der Retrievalqualität und Dokumentenrepräsentation, zu erforschen gibt.

## Literatur

- [Bel61] J. Verhoeff & W. Goffmann & J. Belzer. *Inefficiency of the Use of Boolean Functions for Information Retrieval Systems*. Communications of the ACM 4, S. 557-558, 1961.
- [Fer03] Reginald Ferber. *Information Retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*. dpunkt.verlag Heidelberg, 2003.
- [McG83] Gerard Salton & M. J. McGill. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.
- [Sal71] Gerard Salton. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Englewood Cliffs, N.J., Prentice-Hall, 1971.
- [Wik] Wikipedia. Wikipedia. <http://www.wikipedia.de>.