

Evaluierung eines entscheidungstheoretischen Modells zur Datenbankselektion

Norbert Gövert¹

Universität Dortmund

Zusammenfassung

Eines der zentralen Probleme auf dem Gebiet des Information Retrieval in miteinander vernetzten Datenbanken ist die *Datenbankselektion*: Welche Datenbanken sind zu befragen, damit das Informationsbedürfnis eines Benutzers befriedigt werden kann? Ausgehend von einem entscheidungstheoretischen Modell hierfür stellt es sich heraus, daß die Schätzung der Anzahl relevanter Dokumente in einer Datenbank grundlegend für die Datenbankselektion ist. Es werden Verfahren zur Schätzung dieses Parameters angegeben, die in einem weiteren Schritt einer Evaluierung unterzogen werden. Dabei stellt sich heraus, daß die Schätzung mit einfachen Mitteln nicht möglich ist.

Abstract

A central problem in networked information retrieval is *database selection*. Which databases have to be considered to meet the information need of a user? Originating from a decision theoretic model for database selection one finds out that estimating the number of relevant documents in a database is fundamental for database selection. Approaches to estimate this parameter are presented as well as an evaluation of these approaches. It is exposed that such an estimation with simple methods is not possible.

1 Einleitung

Bislang hat sich die Forschung auf dem Gebiet des Information Retrieval hauptsächlich darauf konzentriert, geeignete Methoden zur Suche in einer abgeschlossenen Dokumentenkollektion zu untersuchen. Insbesondere eine explosionsartig wachsende Zahl von Informationsressourcen und Datenbanken im Internet ließ jedoch weitergehende Fragestellungen im IR entstehen bzw. aktuell werden. Eine der zentralen Fragestellungen bei der Untersuchung vernetzter Information-Retrieval-Systeme ist das *Resource-Discovery-Problem*: Dem Informationsnachfrager stehen eine Vielzahl von weltweit verteilten Datenbanken zur Verfügung. Bevor der Benutzer sein Informationsbedürfnis überhaupt spezifizieren kann, muß er erst einmal die Datenbanken suchen (und finden), die eventuell sein Informationsbedürfnis befriedigen können.

¹Email: goevert@ls6.cs.uni-dortmund.de

Evaluierung eines entscheidungstheoretischen Modells zur Datenbankselektion

Ziel der Datenbankselektion in vernetzten Information-Retrieval-Systemen ist es, die Verteiltheit der Retrievalsysteme bezüglich der Retrievalqualität soweit möglich vor dem Benutzer zu verbergen. Im besten Fall erhält der Benutzer den Eindruck, daß er nicht in mehreren miteinander vernetzten Systemen sucht, sondern in einer großen, aus den Dokumenten der Einzelsysteme zusammengemischten Gesamtkollektion.

Ein entscheidungstheoretischer Ansatz zur Datenbankselektion wird in [Führ 97] vorgestellt. Das ihm zugrundeliegende Szenario zeigt Abbildung 1. Der Benutzer erhält über einen *Broker* Zugriff auf verschiedene Datenbanken. Aufgabe des Brokers ist es, mittels der Anfrage des Benutzers aus den Datenbanken diejenigen auszuwählen, die relevante Dokumente enthalten. Diese werden dann vom Broker befragt. Die Datenbanken ermitteln ihr lokales Suchergebnis und leiten es an den Broker weiter. Dieser muß nun die einzelnen Suchergebnisse zu einem Gesamtergebnis zusammenmischen, welches er dem Informationsnachfrager präsentiert.

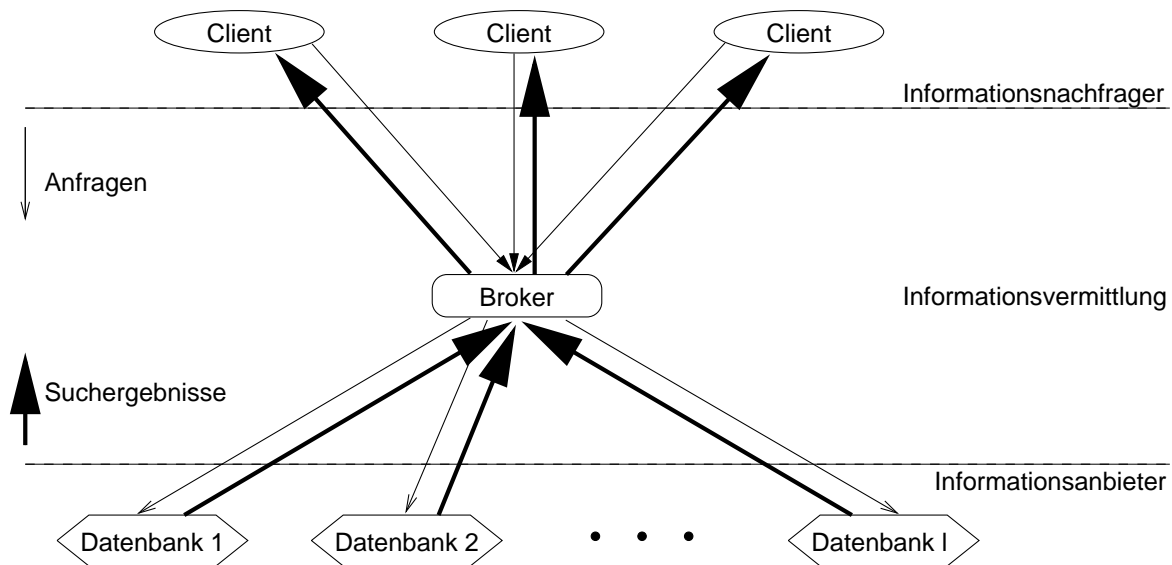


Abbildung 1: Datenbankselektion

Um eine Auswahl von Dokumenten aus den Suchergebnissen der konsultierten Datenbanken vornehmen zu können, muß die Verteilung von relevanten Dokumenten in diesen Datenbanken geschätzt werden. Dazu wird angenommen, daß die beteiligten IR-Systeme optimales Retrieval gemäß des probabilistischen Ranking-Prinzips [Robertson 77] erreichen. D. h. sie ordnen den Dokumenten Relevanzwahrscheinlichkeiten zu und liefern als Suchergebnis eine nach fallenden Retrievalgewichten sortierte Liste von Dokumenten. Betrachtet man nun die Effektivität eines IR-Systems sowie die Anzahl relevanter Dokumente in einer Datenbank, so kann eine Verteilung der relevanten Dokumente im Suchergebnis geschätzt werden.

Innerhalb dieses Artikels wird der Parameter *Anzahl relevanter Dokumente* näher untersucht. Dazu wird in Abschnitt 2 zunächst dargestellt, wie mit Hilfe dieses Parameters eine Verteilung relevanter Dokumente im Suchergebnis einer Datenbank geschätzt werden kann. Basierend auf dem probabilistischen Inferenzmodell für Information Retrieval

[Wong & Yao 95] werden in Abschnitt 3 Verfahren zur Schätzung der Anzahl relevanter Dokumente in einer Datenbank angegeben. Um die Güte dieser Verfahren beurteilen zu können, wurden sie einer Evaluierung unterzogen. Die Ergebnisse der durchgeführten Experimente werden in Abschnitt 4 vorgestellt. Abschnitt 5 gibt eine Zusammenfassung und stellt die aus den Ergebnissen zu ziehenden Konsequenzen dar.

2 Ein Modell für Datenbankselektion

Der in [Fuhr 97] vorgestellte entscheidungstheoretische Ansatz zur Datenbankselektion geht von dem in Abbildung 1 gezeigten Szenario aus: Ein Broker besitzt Zugriff auf eine Menge von l Datenbanken DB_1, \dots, DB_l . Er soll in bezug auf eine Benutzeranfrage eine Auswahl der Datenbanken vornehmen, die relevante Dokumente zum Suchergebnis beisteuern können. Über die reine Datenbankauswahl hinaus soll geschätzt werden, wieviele Dokumente aus den Suchergebnissen der einzelnen Datenbanken abgerufen werden sollen, damit ein optimales Retrievalergebnis erreicht wird.

Neben dem Auswahlkriterium *Relevanz* wird das Kriterium *Kosten* berücksichtigt. Für eine Datenbank DB_i werden Kosten sowohl für die Anfrageprozessierung als auch für das Retrieval von Dokumenten angenommen. Dabei bezeichne c_i^0 die Kosten für die Prozessierung einer Anfrage und c_i^d die Kosten für das Retrieval eines Dokumentes. Nimmt man an, daß die Kosten c_i^d unabhängig vom jeweiligen Dokument sowie von der Anzahl abzurufender Dokumente sind, so ergeben sich für das Retrieval von k Dokumenten ($k > 0$) aus Datenbank DB_i die Kosten zu $c_i(k) = c_i^0 + k \cdot c_i^d$.

Weiterhin werden – wie im Information Retrieval üblich – benutzerspezifische Kosten für das Lesen von relevanten bzw. nicht relevanten Dokumenten angenommen. Betrachtet man den mit dem Lesen eines Dokumentes verursachten Aufwand und den dadurch entstehenden Nutzen, so läßt sich folgern, daß nicht relevante Dokumente höhere Kosten als relevante Dokumente verursachen. Diese benutzerspezifischen Kosten sollen mit C^N für ein nicht relevantes und C^R für ein relevantes Dokument bezeichnet werden.

Neben dem eigentlichen Auffinden relevanter Dokumente kommt es also auch darauf an, dieses mit möglichst geringem Aufwand zu tun. Für den Broker ergibt sich somit folgende erweiterte Optimierungsaufgabe: *Liefere zu einer Anfrage möglichst viele relevante Dokumente zu minimalen Kosten*. Aus dieser Aufgabenstellung lassen sich verschiedene Benutzerstandpunkte für die Datenbankselektion ableiten:

Kriterium Anzahl Dokumente Der Benutzer spezifiziert die Anzahl der Dokumente, die der Broker für ihn ausgeben soll. Hierbei sind zwei Optimierungskriterien denkbar. Es kann sowohl nach minimalen Kosten als auch nach maximaler Anzahl relevanter Dokumente optimiert werden. Im ersten Fall gäbe es jedoch keinen Bezug mehr zur Relevanz von Dokumenten, womit das Kriterium *minimale Kosten* als wenig sinnvoll bezeichnet werden kann. Das Optimierungskriterium soll hier also *maximale Anzahl relevanter Dokumente* heißen.

Kriterium Anzahl relevanter Dokumente Der Benutzer spezifiziert die Anzahl der relevanten Dokumente, die er sehen will. Aufgabe des Brokers ist es, diese Dokumente bei möglichst geringen Kosten anzuzeigen.

Kriterium Kosten Der Benutzer spezifiziert die maximalen Kosten, die entstehen dürfen. Der Broker muß dann möglichst viele relevante Dokumente zu diesen Kosten anzeigen.

Im folgenden wird für den Standpunkt *Kriterium Anzahl relevanter Dokumente* eine Funktion hergeleitet, die die Kosten für eine vom Benutzer gewünschte Anzahl relevanter Dokumente minimiert. Optimierungsfunktionen für die beiden anderen Benutzerstandpunkte sind in [Gövert 97] angegeben.

Um eine Kostenminimierungsfunktion herleiten zu können, sind datenbankspezifische Funktionen notwendig, die die Kosten für das Retrieval einer bestimmten Zahl relevanter Dokumente schätzen. Kann geschätzt werden, wieviele Dokumente insgesamt abgerufen werden müssen, damit man die gewünschte Zahl relevanter Dokumente erhält, so lassen sich die Kosten gemäß des oben angegebenen Kostenmodells berechnen. Wie soll jedoch die Anzahl abzurufender Dokumente geschätzt werden? Dazu sollen für eine zu betrachtende Datenbank DB_i zunächst folgende Parameter als bekannt vorausgesetzt werden:

- Die erwartete Anzahl relevanter Dokumente R_i in Datenbank DB_i und
- die für die Anfrage erwartete Effektivität des betrachteten IR-Systems in Form einer Recall-Precision-Funktion $P_i(R)$.

Die Funktion $s_i(r_i)$, die schätzt, wieviele Dokumente vom Suchergebnis abgerufen werden müssen, um r_i relevante Dokumente zu erhalten, läßt sich nun durch den Zusammenhang zwischen Recall-Precision-Funktion und der Anzahl relevanter Dokumente angeben. Sei s die Zahl der abgerufenen Dokumente. Der Quotient $\frac{r_i}{s}$ gibt die Precision für den Recall-Punkt $\frac{r_i}{R_i}$ an: $\frac{r_i}{s} = P_i(R) = P_i(\frac{r_i}{R_i})$. Durch Umrechnen ergibt sich die gesuchte Funktion. Da immer nur „ganze“ Dokumente abgerufen werden können, findet eine Aufrundung statt:

$$s_i : [0, R_i] \rightarrow \mathcal{N}$$
$$r_i \mapsto s_i(r_i) = \left\lceil \frac{r_i}{P_i(\frac{r_i}{R_i})} \right\rceil$$

Mit Hilfe des Kostenmodells und der Funktion s_i läßt sich nun auch eine Kostenfunktion C_i^r für das Retrieval von r_i relevanten Dokumenten aus Datenbank DB_i angeben. Als Ergebnis werden die geschätzten Gesamtkosten der insgesamt abzurufenden Dokumente geliefert. Die Funktion liefert unendliche Kosten für $r_i > R_i$:

$$\begin{aligned} C_i^r : [0, R_i] &\rightarrow \mathcal{N} \times \mathcal{N} \\ r_i &\mapsto C_i^r(r_i) \end{aligned}$$

mit

$$C_i^r(r_i) = \begin{cases} 0 & : r_i = 0, \\ c_i^0 + s_i(r_i) \cdot c_i^d + r_i \cdot C^R + (s_i(r_i) - r_i) \cdot C^N & : 0 < r_i \leq R_i, \\ \infty & : \text{sonst.} \end{cases}$$

Die Zahl abzurufender Dokumente kann mit $s_i(r_i)$ ermittelt werden. Mit Hilfe der datenbankspezifischen Kostenfunktionen kann nun die globale Kostenminimierungsfunktion für das Retrieval von r relevanten Dokumenten angegeben werden. Das Argument r_i ($i \in \{1, \dots, l\}$) gibt die Anzahl relevanter Dokumente wieder, die der Benutzer aus Datenbank DB_i erhält:

$$\text{Minimiere } C(r_1, \dots, r_l) = \sum_{i=1}^l C_i(r_i) \text{ unter der Nebenbedingung } \sum_{i=1}^l r_i \geq r.$$

Algorithmen zur Berechnung dieser sowie der zu den anderen Benutzerstandpunkten gehörenden Optimierungsfunktionen sind in [Fuhr 97] und [Gövert 97] angegeben.

Grundlegend für die Anwendung des hier vorgestellten Ansatzes zur Datenbankselektion sind also zum einen das Schätzen der Anzahl relevanter Dokumente in einer Datenbank und zum anderen das Schätzen der Effektivität des betrachteten Retrievalsystems. Im folgenden wird der Parameter *Anzahl relevanter Dokumente* näher untersucht. Dazu werden Schätzverfahren angegeben, die später einer Evaluierung unterzogen werden.

3 Schätzung der Anzahl relevanter Dokumente

Ziel dieses Abschnittes ist es, die Anzahl relevanter Dokumente bezüglich einer vorgegebenen Anfrage in einer Datenbank DB mit den Dokumenten $D = \{d_1, \dots, d_{|D|}\}$ zu schätzen. Wie in [Gövert 97] gezeigt wird, ergibt sich der Erwartungswert hierfür aus der Summe der Relevanzwahrscheinlichkeiten für die Dokumente der betrachteten Kollektion:

$$E[NR] = \sum_{d \in D} P(R|q, d) \tag{1}$$

Somit bleiben also die Relevanzwahrscheinlichkeiten für die einzelnen Dokumente zu schätzen. Dazu wird das *probabilistische Inferenzmodell für Information Retrieval* [Wong & Yao 95] herangezogen. Ausgehend von einem Konzeptraum, der das in den Kollektionen vorhandene Wissen repräsentiert, werden drei Maße zur Abschätzung der Relevanz eines Dokumentes bezüglich einer Anfrage angegeben:

Precision-orientiertes Maß: $P(R|q, d) = P(d \rightarrow q) := P(q|d)$

Recall-orientiertes Maß: $P(R|q, d) = P(d \leftarrow q) := P(d|q)$

Ausgeglichenes Maß: $P(R|q, d) = P(d \leftrightarrow q) := \frac{P(q \cap d)}{P(q \cup d)}$

Hier soll der Konzeptraum als Basiskonzepte die in den betrachteten Kollektionen vorkommenden Terme enthalten. Dokumente und Anfragen werden dann als Teilmengen des so entstandenen Konzeptraumes interpretiert werden. Auf dem Konzeptraum wird eine Wahrscheinlichkeitsverteilung definiert. Wird angenommen, daß die Basiskonzepte paarweise disjunkt sind, so ergibt sich die Summe der den Basiskonzepten zugeordneten Wahrscheinlichkeiten zu 1. Für die gesuchte Wahrscheinlichkeitsverteilung soll nun eine *idf*-Gewichtung verwendet werden. Die *inverse Dokumenthäufigkeit* eines Terms t ergibt sich zu $idf(t) = -\log \frac{n_t}{|D|}$, wobei n_t die Anzahl der Dokumente angibt, in denen t vorkommt, und $|D|$ die Gesamtzahl der betrachteten Dokumente bezeichnet. Damit sich die Termwahrscheinlichkeiten insgesamt zu 1 addieren, muß mittels der Summe der *idf*-Gewichte normiert werden. Die Wahrscheinlichkeit eines Terms t ergibt sich somit zu:

$$P(t) = \frac{idf(t)}{\sum_{t \in U} idf(t)} \quad (2)$$

Legt man eine binäre Indexierung für die Repräsentation von Dokumenten und Anfragen zugrunde, so ergeben sich die Wahrscheinlichkeiten von Dokumenten $P(d)$ und Anfragen $P(q)$ durch Aufsummieren der Wahrscheinlichkeiten der in ihnen enthaltenen Terme. Weiterhin können nun bedingte Wahrscheinlichkeiten der Form $P(d|t) = \frac{P(d \cap t)}{P(t)}$ bzw. $P(q|t) = \frac{P(q \cap t)}{P(t)}$ angegeben werden. Sie ergeben sich jeweils zu 1, wenn der Term t im Dokument d bzw. in der Anfrage q enthalten ist, sonst zu 0.

Mit Hilfe der *Tree-Dependence-Assumption* [Wong & Yao 95] erhält man nun eine Interpretation der drei oben angegebenen Maße, die das Vektorraummodell als Spezialfall des probabilistischen Inferenzmodells darstellen:

Precision-orientiertes Maß: $P(q|d) \approx \frac{1}{P(d)} \cdot \sum_{t \in q \cap d} P(t)$

Recall-orientiertes Maß: $P(d|q) \approx \frac{1}{P(q)} \cdot \sum_{t \in q \cap d} P(t)$

Ausgeglichenes Maß: $P(R|q, d) \approx \frac{\sum_{t \in q \cap d} P(t)}{P(q) + P(d) - \sum_{t \in q \cap d} P(t)}$

Diese Werte können nun jeweils für die Relevanzwahrscheinlichkeit $P(R|q, d)$ in den in Gleichung 1 angegebenen Erwartungswert eingesetzt werden. Setzt man zusätzlich die *idf*-basierten Termwahrscheinlichkeiten ein, so ergeben sich nach einigen Umrechnungen die folgenden drei Maße für die erwartete Anzahl relevanter Dokumente in einer Datenbank *DB*:

Precision-orientiertes Maß: $E[NR] \approx \sum_{d \in D} \left(\frac{1}{\sum_{t \in d} idf(t)} \cdot \sum_{t \in q \cap d} idf(t) \right)$

Recall-orientiertes Maß:

$$E[NR] \approx \frac{1}{\sum_{t \in q} idf(t)} \cdot \sum_{d \in D} \sum_{t \in q \cap d} idf(t) = \frac{1}{\sum_{t \in q} idf(t)} \cdot \sum_{t \in q} n_t \cdot idf(t)$$

Ausgeglichenes Maß: $E[NR] \approx \sum_{d \in D} \frac{\sum_{t \in q \cap d} idf(t)}{\sum_{t \in q} idf(t) + \sum_{t \in d} idf(t) - \sum_{t \in q \cap d} idf(t)}$

Anhand der Berechnungsvorschriften für die unterschiedlichen Maße läßt sich nun auch eine Aussage über die zur Berechnung notwendigen Voraussetzungen machen. Zunächst muß der Konzeptraum, für den die Maße berechnet werden sollen, bekannt sein. D. h. die Vorkommenshäufigkeiten der Dokumentterme müssen gegeben sein. Weiterhin ist für die Berechnungen des *Precision-orientierten* und des *ausgeglichenen Maßes* die Betrachtung eines jeden Dokumentes einzeln notwendig, um die Normierung über die Dokumentwahrscheinlichkeit $P(d)$ durchführen zu können. Hingegen brauchen bei der Berechnung des *Recall-orientierten* Maßes nur die Häufigkeiten der Anfrageterme in der betrachteten Datenbank bekannt sein. Somit erweisen sich die Schätzung der Anzahl relevanter Dokumente für das *Precision-orientierte* und das *ausgeglichene Maß* als ungleich aufwendiger als für das *Recall-orientierte* Maß.

4 Experimente

Zur Evaluierung der in Abschnitt 3 hergeleiteten Maße wurden die CACM-Kollektion² mit Relevanzurteilen zu 64 vorgegebenen Anfragen sowie Teile der TREC-Kollektion³ [Harman 95] mit 150 vorgegebenen Anfragen untersucht. Aus der TREC-Kollektion wurden die AP'89-Kollektion (Artikel aus der Tageszeitung *AP-Newswire* vom Jahr 1989) und den auf der zweiten CD-ROM befindlichen Teil der Ziff-Kollektion (Artikel von den *Computer-Select-Disks*, Ziff-Davis-Publishing) ausgewählt. Daneben wurde die Vereinigung der AP'89- und Ziff-2-Kollektion (AP'89+Ziff-2) betrachtet. In Tabelle 1 sind einige Eigenschaften der untersuchten Kollektionen zusammengestellt.

Um einen Eindruck zu gewinnen, wie effektiv die im vorhergehenden Abschnitt beschriebenen Relevanzmaße angewendet auf die hier untersuchten Kollektionen sind, wurden die in Abbildung 2 gezeigten Recall-Precision-Graphen erstellt. Sie zeigen insgesamt keine allzugute Retrievalqualität, wenn man mal die Recall-Precision-Graphen der innerhalb der TREC-Konferenzserie evaluierten Systeme zum Vergleich heranzieht. Andererseits ließen die einfache Form der Dokument- und Anfragerepräsentation sowie die einfachen Retrievalverfahren auch keine besonders guten Retrievalergebnisse erwarten.

Im Mittelpunkt steht nun die Frage, inwiefern mittels der drei Maße die erwartete Anzahl relevanter Dokumente $E[NR]$ geschätzt werden kann. Dazu wurde untersucht, ob eine Kor-

²<ftp://ftp.cs.cornell.edu/pub/smart/cacm/>

³<http://www-nlpir.nist.gov/TREC/>

Evaluierung eines entscheidungstheoretischen Modells zur Datenbankselektion

	CACM	AP '89	Ziff-2	AP '89 + Ziff-2
Anzahl Dokumente	3.204	84.678	56.920	141.598
Anzahl Anfragen	64	150	150	150
Anzahl relevanter Dokumente zu den Anfragen	796	7.312	6.280	13.592
⊙ Anzahl relevanter Dokumente pro Anfrage	12	48	41	90
⊙ Anzahl Terme pro Dokument	23,5	152,4	111,8	136,1

Tabelle 1: Statistische Angaben zu den Testkollektionen

relation zwischen den tatsächlichen Anzahlen relevanter Dokumente NR und den gemäß Abschnitt 3 berechneten Erwartungswerten nachweisbar ist. Die tatsächlichen Anzahlen relevanter Dokumente wurden aus den zu den Kollektionen vorliegenden Relevanzurteilen gewonnen. Anschließend wurde auf die so entstehenden Stichproben für die einzelnen Kollektionen und Verfahren Korrelationsanalysen durchgeführt. Die in Tabelle 2 gezeigten Korrelationskoeffizienten legen die Vermutung nahe, daß im allgemeinen auf diese einfache Weise die Anzahl relevanter Dokumente nicht geschätzt werden kann. Lediglich für die Ziff-2-Kollektion konnten vergleichsweise starke Korrelationen nachgewiesen werden.

	CACM	AP '89	Ziff-2	AP '89 + Ziff-2
Precision-orientiertes Maß	0,129	0,114	0,38	0,143
Recall-orientiertes Maß	0,25	0,08	0,70	0,33
ausgeglichenes Maß	0,22	0,137	0,625	0,241

Tabelle 2: Korrelationskoeffizienten ρ bei Betrachtung der Größen $E[NR]$ und NR

Um weitergehende Evaluierungen durchführen zu können, sind Modifikationen der Schätzverfahren unerlässlich. Dazu soll noch einmal die Definition der Relevanzmaße betrachtet werden. Bislang wurde angenommen, daß $P(R|q, d) = P(q \bowtie d)$ ist ($\bowtie \in \{\leftarrow, \rightarrow, \leftrightarrow\}$). Tatsächlich ist jedoch von folgender Gleichung [Rijsbergen 89] auszugehen:

$$P(R|q, d) = P(R|d \bowtie q)P(d \bowtie q) + P(R|\overline{d \bowtie q})P(\overline{d \bowtie q}) \quad (3)$$

Benötigt wird also eine Schätzung der Wahrscheinlichkeit, daß ein Dokument bezüglich einer Anfrage relevant ist unter der Bedingung, daß die Implikation $\bowtie \in \{\leftarrow, \rightarrow, \leftrightarrow\}$ gilt bzw. nicht gilt. Um eine einfache Betrachtung dieser Wahrscheinlichkeit vornehmen zu können, wird im folgenden der zweite Summand vernachlässigt. Für den Erwartungswert der Anzahl relevanter Dokumente ergibt sich nunmehr:

$$E[NR] \approx \sum_{d \in D} P(R|d \bowtie q)P(d \bowtie q) \quad (4)$$

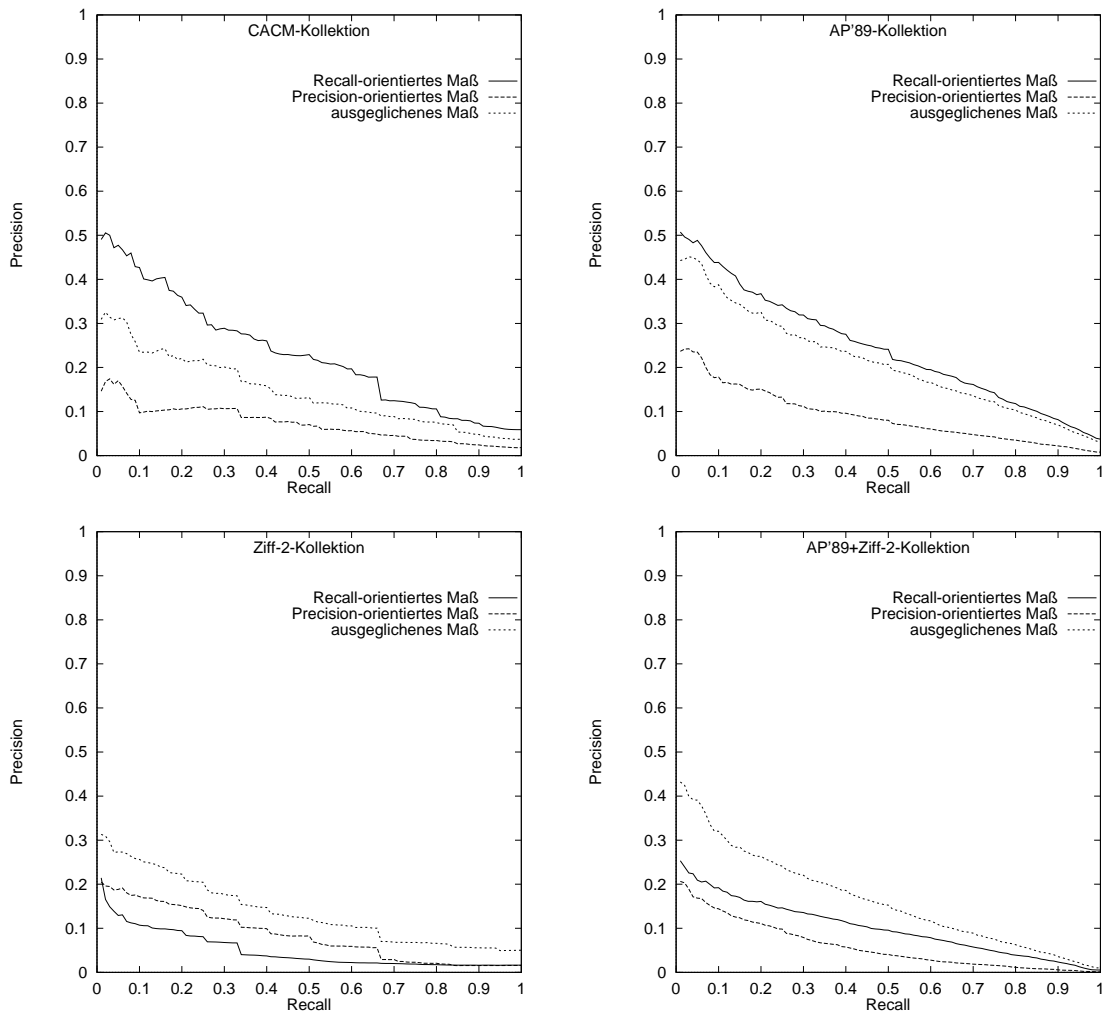


Abbildung 2: Recall-Precision-Graphen

Da die Wahrscheinlichkeit $P(R|d \bowtie q)$ unabhängig vom Dokument d ist, kann der Parameter im Erwartungswert vor die Summe gezogen werden:

$$E[NR] \approx P(R|d \bowtie q) \sum_{d \in D} P(d \bowtie q) \quad (5)$$

$$\implies P(R|d \bowtie q) \approx \frac{E[NR]}{\sum_{d \in D} P(d \bowtie q)} \quad (6)$$

Somit kann $P(R|d \bowtie q)$ auch als unabhängig von der betrachteten Kollektion angenommen werden. In Hinblick auf die erwartete Anzahl relevanter Dokumente lassen sich ohne Schätzung dieser Wahrscheinlichkeit qualitative Aussagen der Form *wahrscheinlich enthält Datenbank DB_1 mehr relevante Dokumente als Datenbank DB_2* oder *die Datenbank DB_1 enthält ca. x -mal sovielen relevanten Dokumenten wie Datenbank DB_2* machen. Das sieht man

Evaluierung eines entscheidungstheoretischen Modells zur Datenbankselektion

bei der Betrachtung der entsprechenden Erwartungswerte für zwei Datenbanken DB_1 und DB_2 . Es ergibt sich mit $P(R|q \bowtie d, DB_1) \approx P(R|q \bowtie d, DB_2)$ nach obiger Gleichung

$$\frac{E_{DB_1}[NR]}{\sum_{d \in DB_1} P(d \bowtie q)} \approx \frac{E_{DB_2}[NR]}{\sum_{d \in DB_2} P(d \bowtie q)} \quad (7)$$

$$\Leftrightarrow \frac{E_{DB_1}[NR]}{E_{DB_2}[NR]} \approx \frac{\sum_{d \in DB_1} P(d \bowtie q)}{\sum_{d \in DB_2} P(d \bowtie q)}, \quad (8)$$

womit solche Aussagen getroffen werden können.

Um nun feststellen zu können, ob eine korrekte Schätzung von $P(R|d \bowtie q, d \in D)$ zu besseren Schätzungen der Anzahlen relevanter Dokumente führt, muß nachgewiesen werden, daß der Parameter für die in Abschnitt 3 eingeführten Verfahren unabhängig von der betrachteten Kollektion ist. Dazu werden die einzelnen Datenbanken in jeweils zwei Subkollektionen aufgeteilt. Darauf wird der Parameter für alle zur Verfügung stehenden Anfragen gemäß der oben angegebenen Formel für die neu entstandenen disjunkten Dokumentmengen berechnet. An Stelle der Erwartungswerte $E_{DB_i}[NR]$ rückt nun die tatsächliche Anzahl relevanter Dokumente (abgeleitet aus den Relevanzurteilen). Die Summen im Nenner von Formel 7 werden für die jeweils zwei Teile der untersuchten Kollektionen gemäß der im vorhergehenden Abschnitt dargestellten Verfahren berechnet. Eine Korrelationsanalyse soll wiederum die Unabhängigkeitsannahme bestätigen.

Die resultierenden Korrelationskoeffizienten für die einzelnen Verfahren und Kollektionen sind der Tabelle 3 zu entnehmen. Für alle Kollektion außer für die AP '89 + Ziff-2-Kollektion konnten relativ starke stochastische Zusammenhänge festgestellt werden; die Korrelationskoeffizienten sind teilweise deutlich größer als 0,6. Die schlechten Ergebnisse bei der AP '89 + Ziff-2-Kollektion sind wohl auf die Unterschiedlichkeit der Dokumente zurückzuführen. Hier wurde die gesamte Kollektion so geteilt, daß die Dokumente der AP '89-Kollektion in dem einen und die der Ziff-2-Kollektion im anderen Teil zu finden waren. Da jedoch der Nachweis eines stochastischen Zusammenhangs insbesondere für unterschiedliche Teilkollektionen von Interesse ist, sind auch diese Untersuchungsergebnisse insgesamt nicht als zufriedenstellend zu bewerten.

	CACM	AP '89	Ziff-2	AP '89 + Ziff-2
Precision-orientiertes Maß	0,589	0,743	0,904	-0,301
Recall-orientiertes Maß	0,693	0,762	0,916	-0,323
ausgeglichenes Maß	0,55	0,742	0,903	-0,322

Tabelle 3: Korrelationskoeffizienten ρ bei Betrachtung von $P(R|q \bowtie d)$ für zwei Teildatenbanken

5 Zusammenfassung

Grundlegend für das in Abschnitt 2 vorgestellte Modell zur Datenbankselektion ist die Schätzung der Anzahl relevanter Dokumente in einer Datenbank. Basierend auf das probabilistische Inferenzmodell für Information Retrieval wurden zur Schätzung dieses Parameters in Abschnitt 3 drei Berechnungsvorschriften angegeben. Insbesondere das Recall-orientierte Maß für die erwartete Anzahl relevanter Dokumente läßt sich mit einfachen Mitteln berechnen. Letztendlich waren dazu nur die Vorkommenshäufigkeiten der Dokumentterme notwendig. Die Evaluierung dieser Verfahren läßt jedoch vermuten, daß mit derart einfachen Mitteln keine zuverlässigen Schätzungen der Anzahl relevanter Dokumente möglich sind.

Um hier zu besseren Ergebnissen zu gelangen, sind weitergehende Untersuchungen der verwendeten Verfahren notwendig. Eine Ausweitung der Experimente auf weitere TREC-Teilkollektion ist notwendig, um die erzielten Evaluierungsergebnisse entweder zu untermauern oder aber um sie verwerfen zu können. Insbesondere sollten aber die bei der Herleitung der Relevanzmaße gemachten vereinfachenden Annahmen untersucht werden. Dazu gehört unter anderem die Repräsentation von Dokumenten und Anfragen durch eine binäre Indexierung.

Was sich bei der Berechnung der Schätzwerte als Vorteil herausstellt, wirkt sich bei der Qualität der Ergebnisse sicherlich negativ aus: Die hier vorgestellten Verfahren zur Datenbankselektion kommen gänzlich ohne Verwendung von Relevance Feedback aus. Die Verbesserung der Retrievalqualität durch die Verwendung von Relevance-Feedback-Methoden wurde schon vielfach nachgewiesen. Die Arbeit von [Voorhees et al. 95] ist ein Beleg dafür, daß auch im Bereich der Fusion von Retrievalergebnissen mittels Relevance Feedback verbesserte Retrievalqualität zu erreichen ist. In [Fuhr 89] wird gezeigt, daß die Schätzung von Relevanzwahrscheinlichkeiten $P(R|q, d)$ mittels Long-Term-Learning-Methoden möglich ist. Dazu werden aus Relevance-Feedback-Daten optimale polynomielle Retrievalfunktionen abgeleitet. Es ist zu prüfen, inwiefern derart ermittelte Relevanzwahrscheinlichkeiten eine Schätzung der Anzahl relevanter Dokumente in einer Datenbank erlauben.

Literatur

Fuhr, N. (1989). Optimum Polynomial Retrieval Functions Based on the Probability Ranking Principle. *ACM Transactions on Information Systems* 7(3), S. 183–204.

Fuhr, N. (1997). *A Decision-Theoretic Approach to Database Selection in Networked IR*. (Eingereicht zur Veröffentlichung).
<http://ls6.cs.uni-dortmund.de/ir/reports/97/Fuhr-97.html>.

Gövert, N. (1997). *Datenbankselektion in vernetzten Information-Retrieval-Systemen*. Diplomarbeit, Universität Dortmund, Fachbereich Informatik.
<http://ls6-www.cs.uni-dortmund.de/~goevert/diploma/>.

Evaluierung eines entscheidungstheoretischen Modells zur Datenbankselektion

- Harman, D.** (1995). The TREC conferences. In: Kuhlen, R.; Rittberger, M. (Hrsg.): *Hypertext - Information Retrieval - Multimedia, Synergieeffekte elektronischer Informationssysteme*, S. 9–28. Universitätsverlag Konstanz, Konstanz.
- van Rijsbergen, C. J.** (1989). Towards an Information Logic. In: Belkin, N.; van Rijsbergen, C. J. (Hrsg.): *Proceedings of the Twelfth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, S. 77–86. ACM, New York.
- Robertson, S.** (1977). The Probability Ranking Principle in IR. *Journal of Documentation* 33, S. 294–304.
- Voorhees, E.; Gupta, N.; Johnson-Laird, B.** (1995). Learning Collection Fusion Strategies. In: Fox, E.; Ingwersen, P.; Fidel, R. (Hrsg.): *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, S. 172–179. ACM, New York.
- Wong, S.; Yao, Y.** (1995). On Modeling Information Retrieval with Probabilistic Inference. *ACM Transactions on Information Systems* 13(1), S. 38–68.