

IQForCE –
Intelligent **Q**uery (Re-)**F**ormulation
with **C**oncept-based **E**xpansion

Stefan Klink

Dissertation

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich IV der Universität Trier

Berichterstatter: Prof. Dr. Bernd Walter
Prof. Dr. Ralph Bergmann
Dekan: Prof. Dr. Dieter Sadowski

Datum der wissenschaftlichen Aussprache: 14. Februar 2006

Zusammenfassung

Die vorliegende Arbeit befasst sich nicht nur mit dem klassischen Problem des Information Retrievals, also dem Finden von relevanten Informationen zu einer gegebenen Anfrage, sondern sie setzt viel früher an und befasst sich bereits mit dem Formulierungsproblem des Suchenden. Noch bevor die ursprüngliche Anfrage an das Information-Retrieval-System übermittelt wird, versucht der hier entwickelte Ansatz dem Benutzer unter die Arme zu greifen und ihm zu helfen, sein Problem in einer geeigneten Anfrage zu formulieren – ohne vorher eine Anfragesprache mit einer komplizierten Syntax oder Operatoren lernen zu müssen.

Es wird ein neuer Weg zur automatischen Verbesserung von Suchanfragen beschritten, der nicht auf manuellen Thesauri oder rein statistischen Auswertungen beruht, sondern mit Hilfe von kollaborativen Lernmethoden das vorhandene Wissen und den reichen Erfahrungsschatz anderer Benutzer verwendet.

Ein ausführlicher Überblick über die Modelle und den State-of-the-Art des ad-hoc Information Retrievals der letzten 50 Jahre dient als Einstieg in die reichhaltige und komplexe Thematik. Da die hier entwickelten Methoden auf dem Vektorraummodell basieren, wird auf dieses in einem eigenem Kapitel detailliert eingegangen.

In der vorliegenden Arbeit wird zur Erweiterung des ad-hoc Retrievals der Begriff des kollaborativen Information Retrievals definiert und das Gesamtszenarium erläutert. Informationen von Suchprozessen früherer Benutzer werden in term-basierten Konzepten automatisch gelernt, ohne dass der Benutzer involviert und bei seiner Suche gestört wird. Die erforderlichen Daten können alleine durch die Beobachtung des Suchverhaltens des Benutzers ermittelt werden. Die gelernten Konzepte beschreiben die Bedeutung des Terms und drücken damit aus, was ein Benutzer mit seiner kurzen Anfrage meint, wenn er einen entsprechenden Term in seiner Anfrage verwendet. Nach dem Lernen werden die Konzepte zur Umformulierung der Benutzeranfrage eingesetzt, um den aktuellen Benutzer bei seiner Suche nach benötigten Informationen zu unterstützen.

Der zunächst entwickelte Basisalgorithmus wird einerseits systematisch Schritt für Schritt mit weiteren Vorgehensweisen erweitert und andererseits mit etablierten Retrievalmethoden kombiniert. Auf diese Art und Weise kann das vorhandene Verbesserungspotenzial des Basisalgorithmus optimal ausgenutzt werden.

Um dem Problem der Mehrdeutigkeit von Termen zu begegnen, wird eine Qualitätseinschätzung entwickelt, welche eine Aussage darüber macht, wie gut das jeweilige term-basierte Konzept gelernt werden konnte und ob es sich lohnt, dieses zur Umformulierung der Benutzeranfrage zu verwenden bzw. wie stark es in der verbesserten Anfrage gewichtet werden soll. Durch den Einsatz dieser Qualitätseinschätzung werden die besten Ergebnisse erzielt. Es ist damit sogar möglich, die hervorragenden Ergebnisse des Pseudo-Relevance-Feedbacks zu übertreffen.

Die Auswirkungen auf die Retrievalleistung der systematischen Erweiterungen, die Kombinationen mit anderen Retrievalmethoden, sowie die Qualitätseinschätzungen werden detailliert auf 28 unterschiedlichen, international standardisierten Testkollektionen ausgewertet und teilweise speziellen Signifikanztests unterzogen.

In einer real existierenden Internet-Suchmaschine beweist der hier entwickelte Ansatz seine Qualitäten und verbessert auch in dieser Anwendung die Retrievalergebnisse des Benutzers durch eine Erweiterung der Anfrage mit term-basierten Konzepten.

Auch bei Agentensystemen und im Peer-to-Peer-Retrieval helfen term-basierte Konzepte individuelle Informationsbedürfnisse von Benutzern schneller und erfolgreicher zu befriedigen. Durch die Umformulierung der ursprünglichen Benutzeranfrage mit den gelernten Konzepten wird sie genauer spezifiziert und der Benutzer bekommt genau die Informationen geliefert, die er mit der Angabe von nur wenigen Wörtern gemeint hat.

Der in dieser Arbeit entwickelte Ansatz, term-basierte Konzepte in einem kollaborativen Szenarium zu lernen, beweist sich damit auf vielfältige Weise in unterschiedlichen Systemen und Anwendungsbereichen als eine mächtige und Gewinn bringende Methode, die jedem einzelnen Benutzer dabei hilft, die benötigten relevanten Informationen zu finden.

Inhaltsverzeichnis

1	Einleitung	1
2	Grundlagen	7
2.1	Einführung in das Information Retrieval	9
2.2	Indexierung	11
2.2.1	Manuelle Indexierung	13
2.2.2	Automatische Indexierung	13
2.2.3	Intelligente Agenten	15
2.2.4	Metadaten, RDF und Annotationen	16
2.3	Retrievalmethoden	18
2.4	Modelle des Information Retrievals	20
2.4.1	Klassische Retrieval Modelle	20
2.4.2	Alternative Retrieval Modelle	31
2.5	Aktuelle Systeme	48
2.5.1	Boole'sche Systeme	48
2.5.2	Vektorraum-Systeme	50
2.5.3	N-Gramm-Methoden	53
2.5.4	Bayessche Netzwerke	54
2.5.5	Statistische Neuronale Netzwerke	55
2.5.6	Kontext-basierende Systeme	57
2.5.7	WordNet-Hierarchie Systeme	58
2.5.8	Ontologie-basierende Systeme	59
2.5.9	Dokumentklassifikation	62
2.5.10	Dokumentcluster-Systeme	63

2.5.11	Wissensbasierende Neuronale-Netzwerk-Systeme	65
2.6	Resumee und Analyse der aktuellen Systeme	68
3	Vektorraummodelle	71
3.1	Frühe Verfahren	72
3.2	Klassisches Vektorraummodell	73
3.3	Strategien zur Termgewichtung	76
3.3.1	idf-Gewichtung	76
3.3.2	Probabilistische Gewichtung	77
3.3.3	<i>tf idf</i> -Gewichtung im SMART System	78
3.3.4	Linearkombinationen von Relevanzhinweisen	79
3.3.5	Termgewichtung im Inquery-System	80
3.3.6	Okapi-Strategie	81
3.4	Klassifikation im Vektorraummodell	82
3.5	Ähnlichkeitsmaße	83
3.5.1	Mathematische Grundlagen	84
3.5.2	Ähnlichkeitsmaße im Überblick	85
3.6	Normalisierte Vektorräume	88
3.7	Relevance Feedback Techniken	89
3.7.1	Taxonomie der Relevance Feedback Techniken	89
3.7.2	Veränderung des Dokumentkorpus	91
3.7.3	Veränderung der Benutzeranfrage	93
3.8	Diskussion	97
4	Konzept-basierte Anfrageerweiterung	101
4.1	„Am Anfang war das Informationsbedürfnis“	101
4.2	Das Problem der Terminologie	102
4.3	Das Problem des Vergessens	106
4.3.1	Collaborative Information Retrieval	106
4.3.2	Eingeschränktes Szenarium	107
4.4	Anfrageerweiterung mit Term-Konzepten	109
4.4.1	Motivation	109
4.4.2	Die Erweiterung der Benutzeranfrage	110

4.5	Lernen von Konzepten (Basisalgorithmus)	111
4.5.1	Das Lernen	111
4.5.2	Diskussion des Algorithmus	113
4.5.3	Beispiele von Konzepten	113
4.5.4	Ergebnisse	114
4.6	Auswahl früherer Anfragen zum Lernen	116
4.6.1	Auswahl der Anfragen im Basisalgorithmus	116
4.6.2	Diskussion der Eigenschaften der Auswahl	117
4.6.3	Erweiterungen der Anfragenauswahl	118
4.7	Auswahl relevanter Dokumente zum Lernen	123
4.7.1	Auswahl relevanter Dokumente im Basisalgorithmus	123
4.7.2	Diskussion der Eigenschaften der Dokumentauswahl	124
4.7.3	Dokumentauswahl mit Pseudo-Relevance-Feedback	125
4.7.4	Dokumentauswahl mit Cluster-Verfahren	131
4.7.5	Dokumentausschnitte mit Passage-Retrieval	142
4.7.6	Individuelle Gewichtung der Dokumente	143
4.8	Qualitätseinschätzung einzelner Konzepte	146
4.8.1	Diskussion der Qualitätseinschätzung	147
4.8.2	Globale Qualitätseinschätzung „best global omega“	147
4.8.3	Individuell gelernte Qualitätseinschätzung „learned binary omega“	152
4.9	Qualitätseinschätzung einzelner Anfrageterme	156
4.9.1	Motivation	156
4.9.2	Positionelle Qualitätseinschätzung	158
4.9.3	Kullback-Leibler Divergenz	158
4.9.4	Durchschnittliche Ähnlichkeit	160
4.9.5	Diskussion der Termqualität	161
4.10	Kombinationen mit Standardverfahren der Anfrageerweiterung	162
4.10.1	Einfache sequentielle Kombination	162
4.10.2	Additive parallele Kombination	166
4.10.3	Ausschließende parallele Kombination	171
4.10.4	Differenzielle parallele Kombination	174

4.10.5	Diskussion und Vergleich der Kombinationen	175
4.11	Kombinationen mit der Dokumenttransformation	177
4.11.1	Motivation	177
4.11.2	Kombination	178
4.11.3	Ergebnisse	179
4.11.4	Diskussion	179
4.12	Diskussion und Erfahrungen mit Konzepten	181
5	Konzepte in einer Suchmaschine	185
5.1	Kurzer Überblick über bisherige Studien	186
5.2	Benutzerverhalten bei der Suche im Internet	186
5.3	Umformulierungen der Benutzeranfragen	188
5.4	Aufbau einer Suchmaschine	190
5.4.1	Interaktionen eines Benutzers	190
5.4.2	Lernen der Konzepte aus den Logdateien	190
5.4.3	Untere Schranke für Konzeptterme	192
5.4.4	Vorteile des Lernen aus Logdateien	192
5.5	Vektorraumsuche	193
5.5.1	Standard Vektorraumsuche	193
5.5.2	Effiziente Berechnung der Skalarprodukte	195
5.5.3	Kombinierte Vektorraumsuche	197
5.6	Phibot Benutzeroberfläche	198
5.7	Analyse der Logdateien	200
5.8	Beschreibung der Testdaten	203
5.8.1	Anzahl und Aufteilung der Suchanfragen	203
5.8.2	Details zu den Suchanfragen und Titelclicks	204
5.8.3	Dokumentkollektion zum Lernen der Konzepte	204
5.8.4	Gelernte Konzepte zu Suchanfragentermen	205
5.9	Manuelle Evaluation	205
5.9.1	Evaluationsmodul der Benutzeroberfläche	205
5.9.2	Testanfragen zur manuellen Evaluation	207
5.9.3	Nützlichkeitsmaß die für manuelle Evaluation	207
5.9.4	Hypothesentest für die manuelle Evaluation	209

5.10	Automatische Evaluation	211
5.10.1	Vergleichsbasis für die automatische Evaluation	211
5.10.2	Testdaten	211
5.10.3	Vergleich der Suchalgorithmen ohne Anfrageerweiterung	212
5.10.4	Vergleich der Suchalgorithmen mit Anfrageerweiterung	213
5.10.5	Irrelevante Dokumente	215
5.11	Diskussion und Erfahrungen mit der Suchmaschine	217
6	Anwendungsszenarien für term-basierte Konzepte	219
6.1	Personalisierte Konzepte	220
6.2	Konzepthierarchien	220
6.3	Mehrdeutigkeiten	222
6.3.1	Motivation und Problematik	223
6.3.2	Aktuelle Systeme	224
6.3.3	Mehrdeutige term-basierte Konzepte	232
6.4	Agentensysteme und P2P-Retrieval	235
6.4.1	Visionen und Ziele	236
6.4.2	Peer-to-Peer Systeme	237
6.4.3	Präzise Befriedigung von Informationsbedürfnisse in interaktiven Arbeitsplätzen	240
6.4.4	Vom Informationsbedürfnis zur Information	241
7	Schlussfolgerung und Diskussion	249
A	Evaluation	253
A.1	Evaluationsmaße	254
A.1.1	Recall und Precision	254
A.1.2	Harmonisches Mittel (F-Maß)	258
A.1.3	E-Maß	259
A.1.4	Durchschnittliche Präzision	259
A.2	Nützlichkeitsmaß	259
A.3	Signifikanztests	262
A.4	Referenzkollektionen	264
A.4.1	Die TREC-Kollektion	265
A.4.2	Andere Kollektionen	268

B	Aufbau der Experimente	271
B.1	Testkollektionen	271
B.2	Aufbereitung der Kollektionen	272
B.2.1	Indexierung	272
B.2.2	Gewichtung der Dokument- und Anfrageterme	273
B.2.3	Transformation in MatLab®	273
B.3	Optimierung von Parametern	277
B.3.1	Pseudo-Relevance-Feedback	278
B.4	Signifikanztests der erzielten Ergebnisse	279
C	Experimente zum Basisalgorithmus	281
C.1	Überprüfung der Grundannahme	281
C.1.1	Variation über die Anzahl der relevanten Dokumente	282
C.1.2	Variation der Anzahl der Terme	285
C.2	Ergebnisse des Basisalgorithmus	286
C.3	Signifikanztests	288
C.4	Ergebnisse der Kombinationen mit Standardverfahren	290
C.5	Ergebnisse der Dokumentauswahl	294
C.5.1	Dokumentauswahl mit Pseudo-Relevance-Feedback	294
C.5.2	Dokumentauswahl mit Cluster-Verfahren	296
C.6	Ergebnisse der Qualitätseinschätzung einzelner Konzepte	300
C.6.1	Globale Qualitätseinschätzung „best global omega“	300
C.6.2	Individuell gelernte Qualitätseinschätzung „learned binary omega“	302
C.7	Ergebnisse einer Kombination mit der Dokumenttransformation	304
	Literaturverzeichnis	307

Abbildungsverzeichnis

2.1	Taxonomie/Schematik der Information Retrieval Modelle.	20
2.2	Drei Kombinationsmöglichkeiten als Venn-Diagramme	22
2.3	Venn-Diagramm einer Kollektion mit dem Anfrageterm „Tisch“	26
2.4	Definition der Wahrscheinlichkeiten	29
2.5	Singular Value Decomposition der Originalmatrix X	36
2.6	Approximation der Originalmatrix durch \hat{X}	36
2.7	Einfaches Bayes'sches Netzwerk	39
2.8	Der CBR-Zyklus nach Aamodt und Plaza	44
2.9	Fälle und Anfrage im strukturellen CBR	46
2.10	Beispiel eines Falls im textuellen CBR	46
2.11	Das INQUERY Bayessche Netzwerk	54
3.1	Anfrage- und Dokumentrepräsentation im Vektorraummodell	74
3.2	Normalisierte Vektoren auf einer Hypersphäre	75
3.3	<i>idf</i> -Gewichtung	77
3.4	Binäre Unabhängigkeitsgewichtung	77
3.5	Original <i>tf idf</i> -Gewichtung mit Kosinus-Normalisierung <i>tfc.tfc</i>	78
3.6	<i>tfc.nfc</i> Termgewichtung	79
3.7	<i>lnc.ltc</i> Termgewichtung	79
3.8	<i>lnu.ltu</i> Termgewichtung	80
3.9	Inquery-Termgewichtung	81
3.10	Okapi BM25 Termgewichtung	82
3.11	Eine Taxonomie für Relevance Feedback Techniken	90
4.1	Vom Problem zur Anfragenrepräsentation	103
4.2	Szenarium des allgemeinen CIR	106

4.3	Eingeschränktes CIR Szenarium	108
4.4	Basisalgorithmus zum Lernen von Konzepten	112
4.5	Ergebnisse mit der CACM-Kollektion	114
4.6	Ergebnisse mit der CRAN-Kollektion	115
4.7	Auswahl früherer Anfragen zum Lernen	117
4.8	Auswahl relevanter Dokumente zum Lernen	123
4.9	Dokumenta Auswahl mit Pseudo-Relevance-Feedback	125
4.10	Vergleichsergebnisse mit der cr-narr-Kollektion	127
4.11	Vergleichsergebnisse mit der CRAN-Kollektion	128
4.12	Dokumenta Auswahl mit dem Cluster-Verfahren	132
4.13	Dendrogramm einer Dokumentkollektion	134
4.14	Dendrogramm mit einer Schranke bei 0,6	137
4.15	'Max-Gap'-Schnitt in einem Dendrogramm	138
4.16	Ergebnisse mit der cr-Kollektion	140
4.17	Ergebnisse mit der med-Kollektion	141
4.18	Qualitätseinschätzung einzelner Konzepte	146
4.19	Einfluss von ω bei der ISI-Kollektion	149
4.20	Erhebliche Verbesserungen auch auf der cran-Kollektion	150
4.21	Verbesserung auf mehr als das Doppelte bei der CR-Kollektion mit 'narrative'-Anfragen	151
4.22	Erhebliche Verbesserungen durch individuelle Qualität	154
4.23	Erhebliche Verbesserungen durch individuelle Qualität	155
4.24	Qualitätseinschätzung einzelner Anfrageterme	157
4.25	Kombinationsergebnisse mit der fr2-desc-Kollektion	164
4.26	Kombinationsergebnisse mit der CACM-Kollektion	165
4.27	Kombinationsergebnisse mit der fr2-desc-Kollektion	167
4.28	Kombinationsergebnisse mit der CACM-Kollektion	168
4.29	Kombinationsergebnisse mit der CACM-Kollektion	169
4.30	Kombinationsergebnisse mit der PT-Kollektion	170
4.31	Verteilung der Termgewichte (entsprechend sortiert)	171
4.32	Kombinationsergebnisse der CACM-Kollektion	172
4.33	Kombinationsergebnisse der CISI-Kollektion	173

4.34	Kombinationsergebnisse der CR-Kollektion	175
4.35	Kombinationsergebnisse der FR89-Kollektion	176
4.36	Kombinationsergebnisse mit der fr2-desc-Kollektion	177
4.37	Kombinationsergebnisse mit FR88-Kollektion	180
4.38	Kombinationsergebnisse der ZF3-Kollektion	181
5.1	Modell der Informationssuche	187
5.2	Interaktion mit Suchmaschinen	188
5.3	Arten der Anfragemodifikation	189
5.4	Ergebnisseite in Phibot	199
5.5	Analyse der Phibot-Logdateien	201
5.6	(r,p)-Diagramm für drei Suchalgorithmen	212
5.7	(r,p)-Diagramm: Expandierte Standardsuche	213
5.8	(r,p)-Diagramm: Expandierte Vektorraumsuche	214
5.9	(r,p)-Diagramm: Expandierte kombinierte Vektorraumsuche	215
6.1	Hierarchie von Konzeptdatenbanken	221
6.2	Auflösen von Term-Mehrdeutigkeiten mit WEBSOM	227
6.3	Grafische Repräsentation eines statischen Bedeutungsvektors	231
6.4	Auswahl eines Vektors mit der passenden Bedeutung	232
6.5	Automatisches Auffinden von Mehrdeutigkeiten mit Cluster	233
6.6	Szenarium in einem Peer-to-Peer IR-System	236
6.7	Suchprozess in einem verteilten IR-System	248
A.1	Relevante Dokumentmengen und Antwortmengen	254
A.2	Beispiel für interpolierte (r,p)-Diagramme	257
A.3	Thema Nr. 165 der TREC-3 Konferenz	267
C.1	Überdeckung der Anfragen	282
C.2	Einfluss der Anzahl der relevanten Dokumente (cacm & CR)	283
C.3	Einfluss der Anzahl der relevanten Dokumente (fr2 & npl)	283
C.4	Einfluss der Anzahl der relevanten Dokumente (cran & med)	284
C.5	Einfluss der Anzahl der relevanten Dokumente (adi & cisi)	284
C.6	Einfluss der Anzahl der Terme (cran & fr2)	285
C.7	Einfluss der Anzahl der Terme (cisi & cacm)	285

Tabellenverzeichnis

2.1	Kanonische Formen von $P(Q T_1, T_2, \dots, T_n)$	40
4.1	Optimale Werte der Qualitätseinschätzung ω	148
5.1	Beispiele für Konzepte für Suchterme	206
5.2	Suchanfragen die für manuelle Evaluation	208
5.3	Nützlichkeitsmaß für die manuelle Evaluation	209
5.4	T-Test für die manuelle Evaluation.	210
5.5	Anzahl positiver und negativer Differenzen	216
A.1	Beispiel für (r,p)-Werte für eine lineare Rangordnung	256
A.2	Kollektionen der TREC-6 (ohne Stopwortreduktion und Stemming).	266
A.3	CACM und ISI und ähnliche Testkollektionen.	270
B.1	Statistische Werte der aufbereiteten Kollektionen	276
B.2	Optimale Werte der Parameter α und θ	278
C.1	Durchschnittliche Präzision des Basisalgorithmus	287
C.2	Ergebnisse der Signifikanztests des Basisalgorithmus	289
C.3	Durchschnittliche Präzision der Kombinationen	293
C.4	Ergebnisse der Dokumentauswahl mit PRF	295
C.5	Dokumentauswahl mit Cluster-Verfahren bei festem Threshold	297
C.6	Ergebnisse der Cluster-Verfahren mit 'Max_Gap'-Strategie	299
C.7	Ergebnisse der globalen Qualitätseinschätzung	301
C.8	Ergebnisse der individuell gelernten Qualitätseinschätzung	303
C.9	Ergebnisse der Kombination mit der Dokumenttransformation	305

*Zwei Drittel dessen, was wir sehen
liegt jenseits unseres Horizonts.*

Chinesisches Sprichwort

Kapitel 1

Einleitung

Die vorliegende Arbeit widmet sich dem grundlegenden Problem des Information Retrievals, welches darin besteht, zu einer von einem Benutzer eines Information-Retrieval-Systems eingegebenen Anfrage genau solche Dokumente aus einer großen und möglicherweise häufig wechselnden Dokumentmenge herauszufinden, welche das Informationsbedürfnis des Benutzers am meisten befriedigen.

Die Informationsflut als Ausgangssituation

Das Suchen in spezifischen Dokumentkollektionen oder im World Wide Web ist eine der aktuellen und weit verbreiteten Instanzen dieses Problems, welches sich in letzter Zeit in zunehmender Weise verstärkt hat. Durch die Benutzung des Computers zur Erstellung und Speicherung von Dokumenten liegen immer mehr Informationen in elektronischer Form vor. Die Einfachheit, Textdokumente elektronisch zu erzeugen und für andere zur Verfügung zu stellen, hat zu einer Potenzierung der für den einzelnen erreichbaren Informationen geführt.

Selbst im akademischen Bereich sprießen neue Konferenzen, Fachzeitschriften und andere Publikationen wie Pilze aus dem Boden und deren Beiträge vergrößern den bereits enormen Datenbestand in einer alarmierenden Weise. [Cleverdon, 1984] schätzte schon vor zwanzig Jahren die Anzahl der jährlichen Veröffentlichungen in den wichtigsten wissenschaftlichen Fachzeitschriften auf ca. 400.000. [Kircz, 1998] stellte fest, dass „*Physics Abstracts*“ alleine in einem Jahr (1996) ca. 174.000 neue Einträge indextierte, wovon ca. 146.500 Artikel in Fachzeitschriften waren.

Schon vor 40 Jahren prophezeiten [Maron und Kuhns, 1960], dass sich die indexierten wissenschaftlichen Informationen etwa alle 12 Jahre verdoppeln und laut einer Studie zum Datenwachstum der Universität von Berkeley verdoppelte sich das Informationsvolumen innerhalb der Jahre 1999 bis 2002 [Lyman *et al.*, 2003]. Aus dieser

Studie geht hervor, dass im Jahr 2002 jeder Mensch im Schnitt 800 Megabyte Daten produziert hat. Das würden etwa zehn Metern Bücherregal entsprechen.

Die Menge an neuen Daten, die auf Papier, Film, sowie optischen und magnetischen Medien gespeichert wurde, hat sich seit 1999 jährlich um 30 Prozent vergrößert und beläuft sich 2002 auf 5.000 Petabyte. Über Telefon, Radio, Fernsehen und Internet wurden 2002 sogar 18.000 Petabyte an Daten ausgetauscht.

Die Studie zum Datenwachstum ergab außerdem, dass der überwiegende Teil der neu generierten Informationen, nämlich 92 Prozent, auf magnetischen Medien wie Festplatten gespeichert wurde. Aufgrund der enormen Datenmengen sollten laut der *EMC Corporation* alle Informationen nicht überall gespeichert werden, sondern zum richtigen Zeitpunkt genau dort verfügbar sein, wo sie benötigt werden.

Laut einer aktuellen Studie der *International Data Corporation (IDC)* wird das Datenvolumen in Unternehmensnetzwerken von 3.200 Petabyte im Jahre 2002 sogar auf ca. 54.000 Petabyte im Jahre 2004 wachsen [Wittenburg und Broeder, 2002].

Diese unvorstellbaren Massen an Daten, denen der Benutzer ausgesetzt ist, macht es ihm enorm schwer, die berühmte „Nadel im Heuhaufen“ zu finden und geradezu unmöglich, alle möglicherweise relevanten Dokumente durchzublättern. Da die Fülle der Informationen kaum mehr überblickt werden kann, besteht zunehmend die Gefahr, Wichtiges zu übersehen oder sich im immer größer werdenden „Heuhaufen“ zu verirren. Diese Situation wird durch die häufig zitierten Begriffe „Informationsflut“ und „Information Overload“ gut charakterisiert: Während früher kaum die Gefahr bestand, etwas zu übersehen, wird man heute häufig von der großen Menge der Informationen derart überflutet, dass man nur noch das berücksichtigen kann, was einem gerade (oft nur zufällig) ins Auge fällt.

Das Problem der Anfrageformulierung

Bevor das World Wide Web oder firmeninterne Intranets mit Unterstützung von Dokument-Management-Systemen auftraten, wurden Information-Retrieval-Systeme hauptsächlich von professionellen Indexierern und Suchspezialisten in Bibliotheken oder Archiven bedient, z. B. für eine gezielte Suche nach benötigter Literatur oder für fachspezifische oder -übergreifende Recherchen. Üblicherweise fungierten diese Suchspezialisten als „Übersetzer“ und versuchten das in einem interaktiven Dialog mit dem Benutzer entwickelte Informationsbedürfnis in eine für das System angepasste Weise in das Retrieval-System als Anfrage einzugeben. Der dabei wichtige Unterschied zwischen professionellen Suchspezialisten und unbedarften Benutzern ist, dass erstere zum einen die Dokumentkollektion und deren interne Repräsentation kennen und zum anderen erfahren sind im Umgang mit Boole'schen Operatoren und der Kombination von Suchbegriffen.

Die heutigen modernen Information-Retrieval-Systeme, wie z. B. Suchmaschinen im World Wide Web oder Dokument-Management-Systeme, sind jedoch direkt für sol-

che unbedarfte Benutzer gedacht – ohne die Unterstützung von Suchspezialisten. Dies führt häufig dazu, dass der Benutzer hilflos im „Heuhaufen herum stochert“ und sich in den gelieferten (meist irrelevanten) Dokumenten verliert:

Ein Benutzer sucht nach etwas ganz bestimmten, weiß aber nicht genau, wie er es formulieren soll. Häufig gibt er dann wahllos irgendwelche Wörter als Anfrage ein, die ihm gerade zu diesem Thema in den Sinn kommen und er lässt sich dann „überraschen“ welche Dokumente geliefert werden. Von diesen Dokumenten lässt er sich dann inspirieren, wie er die ursprüngliche Anfrage umformulieren könnte, um hoffentlich diesmal Dokumente zu erhalten, die ihm weiter helfen.

Hat der Benutzer jedoch in der Initialanfrage schon die falschen Wörter gewählt, so erhält er im schlimmsten Fall irreführende Dokumente, die ihn in eine falsche Richtung leiten und die Anfrage eher verschlimmern als verbessern. Dies hat dann zur Folge, dass die Suche nach den dringend benötigten Informationen gänzlich scheitert.

Zielsetzung der Arbeit

In der vorliegenden Arbeit wird das eingangs erläuterte Problem des Information Retrievals auf das Problem der Anfrageformulierung reduziert, welches mit der Beantwortung der folgenden Fragen für den Benutzer gelöst wird:

1. Wie kann man jenseits von (Ähnlichkeits-) Thesauri, Ontologien und co-occurrence-basierenden Verfahren einem Benutzer eines Information-Retrieval-Systems bei der Formulierung seines Informationsbedürfnisses unterstützen und mit Hilfe von Anfrageerweiterungstechniken die ursprüngliche Benutzeranfrage derart umformulieren, dass die Leistung von Retrievalsystemen nachhaltig verbessert wird?
2. Wie kann man die Anfrageerweiterung automatisieren, um dem Benutzer auf eine unaufdringliche Weise möglichst schnell zu relevanten Dokumenten zu führen – ohne ihn interaktiv, z. B. durch die manuelle Auswahl von Termen, in den Erweiterungsprozess einzubinden und ihn damit bei der schnellen Suche zu behindern?
3. Wie kann man für die obige Methodik das vorhandene Wissen und die Erfahrungen früherer Benutzer oder früherer Suchen desselben Benutzers während der aktuellen Suche Gewinn bringend nutzen?
4. Wie kann die in dieser Arbeit entwickelte Methodik genutzt werden, um Mehrdeutigkeiten einzelner Terme der Benutzeranfrage zu identifizieren und automatisch aufzulösen?

Wissenschaftliche Beiträge

Die wissenschaftlichen Beiträge der vorliegenden Arbeit können in vier Bereiche eingeteilt werden:

1. Entwicklung von kollaborativ gelernten, term-basierten Konzepten zur automatischen Verbesserung von Benutzeranfragen

Im Rahmen dieser Arbeit wurde eine Methode entwickelt, welche das Formulierungsproblem eines Suchenden vermindert. Hierzu wurde der Begriff des kollaborativen Information Retrievals entwickelt. Innerhalb eines eingeführten Mehrbenutzer-Szenariums wurden Konzepte gelernt, welche auf einzelnen Termen der Suchanfrage basieren. Diese wurden dann zur automatischen Verbesserung der Anfrage eingesetzt.

Durch den Einsatz eines kollaborativen Systems ist es möglich, vom Wissen und der Arbeit anderer Benutzer zu profitieren und eine Verbesserung der eigenen aktuellen Suche zu erzielen.

2. Systematische Erweiterung der Basismethode und Analyse von Kombinationen mit Standardverfahren

Die Grundidee der entwickelten Methode wurde auf systematische Weise Schritt für Schritt mit verschiedenen Verfahren erweitert. Hierzu zählen z. B. linguistische Verfahren zur Auswahl ähnlicher Anfragen, Pseudo-Relevance-Feedback und Cluster-Verfahren zur Filterung der „besten“ relevanten Dokumente, sowie die Entwicklung einer Qualitätsabschätzung der gelernten Konzepte und einzelner Anfrageterme. Insbesondere die Qualitätsabschätzung der gelernten Konzepte zeigte auf den verwendeten Testkollektionen ein enormes Verbesserungspotential und erzielte die besten Retrievalergebnisse.

Durch das Sprichwort „Vier Augen sehen mehr als zwei“ motiviert, wurde die entwickelte Methode mit anderen Retrievalverfahren auf verschiedene Weisen kombiniert. Eine Analyse und Vergleich der Resultate bestätigte das Sprichwort auch in diesem Zusammenhang und zeigte eine Verbesserung der Retrievalergebnisse.

3. Verwendung term-basierter Konzepte in existierenden Internet-Suchmaschinen und Peer-to-Peer-Systemen

Die Implementierung term-basierter Konzepte ist sowohl Grundlage für Forschungsarbeiten, welche sich an typische Retrievalprobleme anpassen lassen, als auch für praktisch eingesetzte Systeme, wie z. B. Internet-Suchmaschinen oder Question-Answering-Systeme.

Term-basierte Konzepte dienten auch als Basis einer Diplomarbeit [Henninger, 2002] und wurden am DFKI in verschiedene Forschungs- und Industrieprojekte erfolgreich eingebracht, unter anderem in den Projekten mFacts, Adaptive READ und EPOS.

Mit Hilfe term-basierter Konzepte wurde die Internet-Suchmaschine *phibot* entwickelt, welche die Anfrage eines aktuellen Benutzers automatisch mit vorher gelernen Konzepten erweitert und damit die Retrievalergebnisse nachweislich verbessert. Mit einer Analyse des Klickverhaltens des Benutzers mit Hilfe der gespeicherten Log-Dateien konnten die Konzepte stets aktuell gehalten werden.

Das Forschungsprojekt EPOS zeigt gegenwärtig, dass term-basierte Konzepte auch in Peer-to-Peer-Systemen zur Verbesserung der Suchanfrage und zur Qualitätseinschätzung und -steigerung der Informationssuche eingesetzt werden können.

4. Einsatz von Konzepten zum Aufspüren und Auflösen von Mehrdeutigkeiten

Ein wesentliches Problem bei der Suche nach relevanter Information ist die Mehrdeutigkeit von Termen. Dies führt dazu, dass die Präzision der Suche vermindert wird und Informationen geliefert werden, welche im aktuellen Kontext irreführend sind.

Die vorliegende Arbeit widmet sich diesem Problem und analysiert hierzu bekannte Ansätze zum Aufspüren und Auflösen von Mehrdeutigkeiten. Zur Lösung des Problems werden neue Ansätze entwickelt, welche sowohl zum Aufspüren als auch zum Auflösen von Mehrdeutigkeiten eingesetzt werden können. Als Basis dienen hierzu die im Rahmen dieser Arbeit entwickelten Qualitätsabschätzungen für Konzepte und einzelne Anfrageterme.

Kapitelübersicht

Die vorliegende Arbeit ist wie folgt aufgebaut:

Kapitel 2 *Grundlagen*

Zunächst werden in Kapitel 2 die Grundlagen und der aktuelle Stand der Technik des Information Retrievals beschrieben.

Kapitel 3 *Vektorraummodelle*

Dieses Kapitel gibt eine Einführung in die Vektorraummodelle und erläutert alle wichtigen Aspekte, die im Rahmen dieser Arbeit eingesetzt wurden.

Kapitel 4 *Konzept-basierte Anfrageerweiterung*

Dieses Kapitel stellt das im Rahmen der vorliegenden Arbeit entwickelte Modell zur Umformulierung einer Benutzeranfrage vor und analysiert in einer detaillierten Weise systematische Erweiterungen der Basisidee.

Kapitel 5 *Konzepte in einer Suchmaschine*

Dieses Kapitel analysiert den praktischen Einsatz des entwickelten Modells in einer weltweit öffentlichen Internet-Suchmaschine zum Retrieval von Informationen im Themenbereich der Physik.

Kapitel 6 *Anwendungsszenarien für term-basierte Konzepte*

Dieses Kapitel zeigt mit Hilfe von Anwendungsszenarien, wie man mit den gelernten Konzepten auf spezifische Eigenschaften der Benutzer eingehen kann. Weiterhin werden Ideen für Agentensysteme und für ein verteiltes Peer-to-Peer Information Retrieval diskutiert.

Kapitel 7 *Schlussfolgerung und Diskussion*

Kapitel 7 fasst die erreichten Ergebnisse zusammen und diskutiert die Beiträge der Arbeit.

Anhang A *Evaluation*

Anhang A erläutert die Grundlagen zu den Verfahrensweisen der gemachten Analysen und Evaluationen dieser Arbeit und beschreibt die verwendeten Referenzkollektionen.

Anhang B *Aufbau der Experimente*

Anhang B beschreibt, wie die Referenzkollektionen für die Analysen aufbereitet wurden, und diskutiert die Verwendung und Optimierung von Parametern.

Anhang C *Experimente zum Basisalgorithmus*

Anhang C beschreibt die durchgeführten Analysen und listet tabellarisch alle Ergebnisse über alle Referenzkollektionen auf.