IQForCE – **I**ntelligent **Q**uery (Re-)**For**mulation 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

Abbildungsverzeichnis

Tabellenverzeichnis

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 indexierte, 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 gelernten 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.