Zwischenbericht: Fuzzy-Suchmethodik für einen kooperativen Rechercheassistenten

16. Oktober 2001

1 Projektdaten

Die Universitätsbibliothek Bielefeld und der Lehrstuhl für Technische Informatik der Technischen Fakultät der Universität Bielefeld haben am 13.11.1998 ein Projekt zur Konzeption, Implementierung und Evaluation eines intuitiv bedienbaren Rechercheassistenten für die Literaturrecherche, basierend auf einer neuartigen Fuzzy-Suchmethodik, beantragt. Der Antrag wurde am 8.4.1999 von der DFG genehmigt. Das Projekt, das dem Förderprogramm ,,Modernisierung und Rationalisierung in wissenschaftlichen Bibliotheken" zugeordnet ist, wurde am 1.4.2000 begonnen und läuft bis zum 31.3.2002.

2 Zusammenfassung

Als vorläufiges Projektergebnis kann in diesem Zwischenbericht bereits ein mit allen wesentlichen Funktionalitäten - wie im Projektantrag beschrieben - ausgestatteter Rechercheassistent vorgestellt werden, der nach gegenwärtigem Stand simultane Suchen im OPAC der UB Bielefeld und der Elsevier Science-Aufsatzdatenbank ermöglicht und im Dauerbetrieb für jedermann verfügbar ist.[1] Bei der Entwicklung wurde von Anfang an darauf geachtet, typische Fehler ähnlicher Projekte zu vermeiden. Funktionsumfang und Benutzerschnittstelle des Systems wurden daher im ständigen Austausch zwischen den Entwicklern und den Bibliotheksmitarbeitern konzipiert, um sicherzustellen, daß der Benutzer tatsächlich von praxisrelevanten Vorteilen profitieren kann. Durch Verwendung von Fuzzy-Suchmethoden konnten Retrieval-Qualität und -Komfort so deutlich erhöht werden. Der Einsatz optimierter Anfragebearbeitungsstrategien stellt kurze Antwortzeiten auch bei Suchen in sehr großen Datenbeständen sicher - in Kürze wird das durchsuchbare Volumen durch die Einbindung weiterer Datenbanken auf über 25 Mio. Dokumente erweitert werden.

3 Projektbeschreibung

3.1 Hintergrund

Das Projekt ,,Fuzzy-Suchmethodik für einen kooperativen Rechercheassistenten" ist vor dem Hintergrund einer enormen Zunahme elektronisch verfügbarer bibliotheksrelevanter Metadaten (Katalogdaten, Literaturnachweise) und Dokumente zu sehen, die den Benutzern zunehmend auch über Metasuchen (simultane Suchen über heterogene Datenbanken) verfügbar gemacht werden. Die Zunahme an Quantität

[1] http://www.ub.uni-bielefeld.de/rechercheassistent/

zusammen mit den Abstrichen, die bei Metasuchen i.d.R. im Hinblick auf Recherchefunktionalität und -qualität gemacht werden müssen, führen einerseits zu einer bisher nicht gekannten Informationsfülle (in Bezug auf die bereitgestellten Informationsmengen), andererseits zu einem - zumindest subjektiv wahrgenommenen - Informationsmangel hinsichtlich Zahl und Qualität der zu einer bestimmten Fragestellung vorhandenen relevanten Nachweise. Dieser Situation kann und sollte mit einer verbesserten Retrieval-Methodik, die die Schwächen herkömmlicher Retrieval-Methoden vermeidet und auf den Techniken der Fuzzy-Logik und der Metasuche aufsetzt, begegnet werden.
Ein bedeutender Schwachpunkt herkömmlicher Suchpraxis, auf den bisher nicht genügend hingewiesen wurde, ist der, daß der Benutzer nicht in die Lage versetzt wird, sein terminologisches Wissen über den Recherchegegenstand voll in die Suche einzubringen: Ihm stehen hierfür i.a. eine ganze Reihe von Begriffen (einschließlich Synonyme, Aspektbegriffe etc.) zur Verfügung. Zur Vermeidung von Rechercheaufwand und -problemen beschränkt sich der Durchschnittsbenutzer bei der Suche üblicherweise aber nur auf wenige (zwei bis drei) Suchbegriffe. Er sucht daher in Abhängigkeit vom Sucherfolg häufig iterativ mit unterschiedlichen Suchbegriffskombinationen. Daraus resultieren sich überschneidende Treffermengen; Suchbegriffskombinationen, über die relevante Treffer hätten ermittelt werden können, unterbleiben usw. Im Unterschied dazu ermöglicht der Rechercheassistent problemlos die Verarbeitung vieler Suchbegriffe in einer einzigen Suche, d.h. man gelangt mit der eingesetzten Suchtechnik i.d.R. zu befriedigenden Suchergebnissen, ohne daß man - wie im Falle boolesche Suchen - explizite Suchbegriffsbeziehungen deklarieren und Vermutungen über die Größe von Treffermengen bzw. Teiltreffermengen anstellen muß.
Auch wenn die terminologische Beschreibung des Recherchegegenstandes eindeutig und erschöpfend ist, hat die üblicherweise verwendete boolesche Suchtechnik bekannte Nachteile, die sich zusammenfassend dadurch charakterisieren lassen, daß der Nutzer unzureichende Informationen über die im Gesamtbestand enthaltenen relevanten Dokumente erhält: im Falle einer UND-Verknüpfung der Suchbegriffe dadurch, daß ggf. keine oder zu wenig Treffer - in Bezug auf nicht ermittelte teilrelevante Treffer (die nicht sämtliche Suchbegriffe enthalten) - erzielt werden; im Falle einer ODER-Verknüpfung dadurch, daß die relevanten Treffer oft in einer großen, unsortierten Menge untergehen.
Des weiteren besteht in herkömmlichen Suchsystemen i.a. nicht die Möglichkeit, die Wichtigkeit der einzelnen Suchbegriffe für das Suchergebnis zu spezifizieren.

3.2 Suchmethodik

Die beiden letztgenannten Mängel motivieren den Wunsch nach einer Suchtechnik, deren Ergebnis sich als relevanzsortierte Trefferliste präsentiert, bei der die Relevanz eines Nachweises / Dokuments zum einen durch die Anzahl der in ihm gemeinsam auftretenden Suchbegriffe, und zum anderen durch deren individuelle Wichtigkeit bestimmt wird. (Letztere wird durch einen Gewichtungsfaktor spezifiziert.) In der Formulierung der einschlägigen Theorie geht es um die Realisierung einer graduellen Aggregation über Suchbegriffsmengen mit graduellen Suchkriterien (hier gewichtete Suchbegriffe). Der im Projekt verfolgte Ansatz ist dem Fuzzy-Information-Retrieval, insbesondere der Fuzzy-Aggregationsmethodik, zuzuordnen und basiert auf neueren Arbeiten am Lehrstuhl für Technische Informatik. Der Ansatz ist sowohl theoretisch stimmig als auch rechnerisch effizient, d.h. er gestattet im Zusammenspiel mit geeigneten Realisierungskonzepten eine Anwendung auf operationale Datenbanken (unter Realbedingungen).
Über Fuzzy-Quantoren (auch „Aggregationsoperatoren" genannt), die durch natürlichsprachliche, intuitiv verständliche Quantoren („fast alle", „viele", „einige"

usw.) ausgedrückt werden können, kann de facto spezifiziert werden, welchen Stellenwert für die Relevanzsortierung das gemeinsame Vorkommen von Suchbegriffen im Vergleich zur individuellen Termgewichtung hat: Der (fuzzy-interpretierte) Aggregationsoperator ,,UND" berücksichtigt in maximaler Weise den ersten Aspekt, der (ebenfalls fuzzy-interpretierte) Operator ,,ODER" in maximaler Weise den letzteren Aspekt (der Termgewichtung). Die dazwischenliegenden Aggregationsoperatoren (wie die eingangs genannten) interpolieren zwischen diesen beiden Möglichkeiten.
Das Beispiel aus den Abbildungen 1-4 verdeutlicht dies: Eine Anfrage wie in Abbildung 1 liefert bei der klassischen booleschen Interpretation mittels UND überhaupt keine Ergebnisse (Abb. 2). Eine disjunktive Verknüpfung der Suchbegriffe (boolesches ODER) liefert sehr viele Dokumente, wobei aber die Ergebnispräsentation von Natur aus unstrukturiert ist, und die relevanten Dokumente in der Masse der nicht relevanten Dokumente untergehen (vgl. Abb. 3 - in diesem konkreten Fall sind sogar die ersten 10 gelieferten Dokumente eher nicht relevant!). Bei einer fuzzy-interpretierten Anfrage sind die Ergebnisse nach Relevanz geordnet (siehe Abb. 4) - die für den Benutzer interessantesten Dokumente werden zuerst präsentiert!

Abbildung 1: Suchmaske. Die Suchmaske des Rechercheassistenten bietet umfangreiche Kontrollmöglichkeiten und erlaubt z.B. eine manuelle Gewichtung der Suchbegriffe. Die Voreinstellungen sind jedoch so gewählt, daß auch dann mit sinnvollen Ergebnissen zu rechnen ist, wenn der Benutzer lediglich einige Suchbegriffe eingibt.

Abbildung 2: Ergebnis bei boolescher Aggregation (mittels AND). In einem traditionellen Retrieval-System scheitert die Suche und eine Reformulierung der Anfrage wird erforderlich.

Titel	Jahr	Ranking
Unsicherheit, Unschärfe und rationales Entscheidendie Anwendung von **Fuzzy**-Methoden in der Entscheidungstheorie; mit 14 Tabellen	2001	-
Die grenzenlose Unternehmung **Information**, Organisation und Management; Lehrbuch zur Unternehmensführung im **Informationszeitalter**	2001	-
Do domestic investors have more valuable **information** about individual stocks than foreign investors?	2001	-
Quantum computation and quantum **information**	2001	-
Wissenschaft zwischen **Information** und Geheimhaltungüber einen blinden Fleck in den Lehren zu Art. 5 Abs. 3 GG	2001	-
Special issue: Enhancing organizations 'intellectual bandwidth': the quest for fast and effective value creation...	2001	-
Special issue: Soft computing for medical image processing: selected papers of a special session on Soft Computing in Medical Image Processing of the 5th International Conference on Soft Computing and **Information**/Intelligent Systems (IIZUKA '98) in 1998. Soft computing for medical image processing	2001	-
Special issue on accessing **information** in spoken audio	2001	-
Die Ökonomie der **Information**sgesellschaft	2001	-
Population growth, technological adoption and economic outcomesa theory of cross-country differences for the **information** era	2001	-

Abbildung 3: Die ersten 10 Ergebnisse bei boolescher Anfrage (OR). Die alternative Verknüpfung mit ODER erbringt in traditionellen Retrieval-Systemen ebenfalls nicht das gewünschte Ergebnis und liefert zu geringe Trefferrelevanz.

Titel	Jahr	Ranking
Special issue Uncertainty in geographic **information** systems and spatial data...	2000	0.62
Uncertainty in intelligent and **information** systems	2000	0.62
Organization of multimedia resourcesprinciples and practice of **information retrieval**	1999	0.60
Fuzzy sets in approximate reasoning and **information** systems	1999	0.60
Special issue: From **information retrieval** to knowledge management enabling technologies and best practices...	1999	0.60
Information-Retrieval Einführung in Grundlagen und Methoden; ein Vorlesungsskript	1998	0.58
Information and documentation - **Information retrieval** (Z39.50)...	1998	0.58
Intelligent multimedia **information retrieval**	1997	0.56
Nicht-lineares **Information Retrieval** in der juristischen **Information**ssuche	1997	0.56
Semantische Umfeldsuche im **Information-Retrieval** in Online-Katalogen	1997	0.56

Abbildung 4: Die ersten 10 Ergebnisse bei einer Fuzzy-Anfrage. Die Fuzzy-Verknüpfung im Rechercheassistenten findet auf Anhieb die relevanten Dokumente.

3.3 Weitere konzeptionelle Merkmale

Neben der ausführlich dargestellten Fuzzy-Suchmethodik, die die Retrieval-Qualität vor allem gegenüber dem booleschen Retrieval deutlich erhöht, zeichnet sich der Rechercheassistent vor allem durch seine Konzeption als Meta-Suchmaschine aus. Diese bietet von ihrer Architektur her die Möglichkeit, heterogene Datenbestände einzubinden. Dabei weist sie ein hohes Maß an Skalierbarkeit auf und ist sowohl für große Datenvolumina als auch für hohe Nutzungsfrequenzen geeignet. Die einfache Bedienbarkeit, die durch ein ergonomisches Java-Applet, welches eine dynamische Oberfläche darstellt, erreicht wurde, macht den Rechercheassistenten in Kombination mit der Zugriffsmöglichkeit über das WWW einem großen Nutzerkreis zugänglich.

3.4 Funktionalitäten der Pilotanwendung

Ein diesen Anforderungen entsprechender Rechercheassistent wird seit ca. zwei Wochen als Testanwendung den Benutzern der UB Bielefeld über die Homepage im Dauerbetrieb angeboten, wobei auf das Testangebot auf verschiedene Weise aufmerksam gemacht wird. Der Rechercheassistent ist dabei in ein Menü von erklärenden und kommentierenden Seiten eingebunden. Wie eingangs erwähnt, ermöglicht der Rechercheassistent derzeit über eine einheitliche Such- und Präsentationsschnittstelle eine simultane Suche im OPAC der Bibliothek und der Elsevier Science-Datenbank.
Die Suche arbeitet mit Voreinstellungen für den Aggregationsoperator und die Gewichtungsfaktoren (letztere in aufeinanderfolgenden Eingabezeilen abnehmend). Die Voreinstellungen können über Pull-down-Menü (Aggregationsoperatoren) bzw. Schieberegler (Termgewichte) geändert werden (vgl. auch Abbildung 1). Die standardmäßig angezeigte Suchmaske erlaubt die Eingabe praktisch beliebig vieler Stichwörter und Phrasensuchbegriffe aus den Titeln und (datenbankabhängig) aus dem Abstract- bzw. Deskriptorvokabular. In das Fuzzy Retrieval können darüber hinausgehend (im Antrag nicht erwähnt) bei Bedarf auch Autorennamen und Erscheinungsdaten einbezogen werden - über Suchfenster, die zusätzlich über Mausklick geöffnet werden können.
Die Ergebnisanzeige realisiert relevanzsortierte Trefferseiten (vgl. Abbildung 4) - wahlweise mit Treffern aus beiden Datenbanken oder jeweils einer - unter Angabe jeweils des Ranking (Relevanzfaktors). Die Realisierung der beinahe selbstverständlichen Forderung, dem Benutzer zusätzlich die ihm bekannten Standard-Funktionalitäten von Datenbanken zur Verfügung zu stellen (z.B. Highlighting (z.Z. noch nicht vollständig implementiert), Verzweigung aus den Trefferlisten in jeweils datenbankspezifische Langanzeigen) war im Fall der Funktion des Zurück- (und Vor-) Blätterns nicht-trivial, da die Seiten sukzessiv aufgebaut werden, und es insofern keine fixen Ergebnismengen gibt. (Aus dem gleichen Grunde kann die prinzipiell wünschenswerte Anzeige einer Gesamttrefferzahl nach Absetzen der Suchanfrage nicht realisiert werden.)

4 Bisheriger Projektverlauf / Realisierung

Im bisherigen Projektverlauf konnten die Arbeitspakete plangemäß erfüllt werden. Die Entwicklung lief dabei über einige Arbeitspakete hinweg parallel ab, um möglichst frühzeitig zu einem funktionsfähigen Prototypen zu gelangen. Dadurch wurde sichergestellt, daß potentielle Probleme frühzeitig erkannt und behoben werden konnten, und es konnten dank der engen Kooperation der Technischen Informatik und der Universitätsbibliothek die praxisrelevanten Aspekte berücksichtigt werden, so daß sichergestellt ist, daß der Endbenutzer tatsächlich entscheidenden Nutzen aus dem Rechercheassistenten zieht.

4.1 Architektur

Zur Realisierung des Rechercheassistenten wurde eine 3-Schichten-Architektur gewählt, wie sie in Abbildung 5 zu sehen ist: Benutzerseitig steht eine als Java-Applet implementierte *ergonomische Oberfläche* zur Verfügung (vgl. Abb 1). Diese wurde in ständigem Austausch zwischen Bibliothek und Entwicklern verfeinert, und erlaubt dem Endbenutzer die komfortable Eingabe auch komplexer Anfragen bestehend aus mit Gewichten versehenen Stichworten, Autoren- und Datumseingabefeldern sowie der Möglichkeit, die zu durchsuchenden Datenbanken und andere Rahmenbedingungen wie ,,Timeout" und gewünschter ,,Trefferanzahl pro Ergebnisseite" einzugeben.

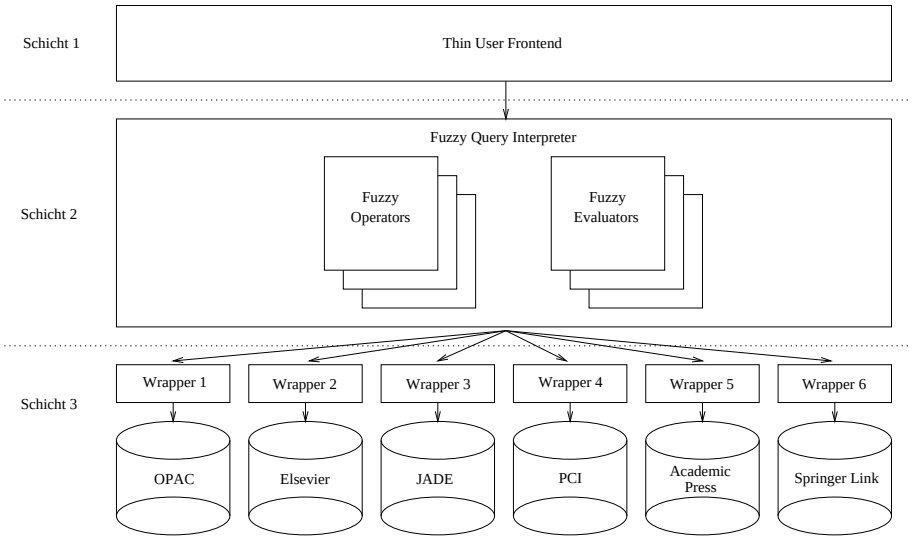

Abbildung 5: Architektur. Benutzerseitig steht ein ergonomisches Java-Applet als Eingabemaske im WWW zur Verfügung. Der Fuzzy-Query-Interpreter ist ein verteiltes, performance-optimiertes System, welches die aufwendige Interpretation und Abarbeitung der Benutzeranfragen übernimmt. Die Anbindung an die Datenbanken geschieht über kompakte Wrapper, durch die der Fuzzy-Query-Interpreter über eine einheitliche Schnittstelle mit den Datenbanken kommunizieren kann.

Die Implementierung als Java-Applet erlaubt eine wesentlich dynamischere Oberfläche als es mit konventionellen, HTML-basierten Eingabeformularen möglich ist. Es stehen wesentlich mehr verschiedene Eingabeelemente, wie z.B. Schieberegler für die Gewichte zur Verfügung, die durch ein einheitliches Look-and-Feel eine einfache Bedienung erlauben. Durch die Dynamik der Oberfläche ist es möglich, eine direkte Eingabeprüfung vorzunehmen, und den Benutzer *unmittelbar* auf mögliche Eingabefehler hinzuweisen.
Der *Fuzzy-Query-Interpreter* ist ein modulares System, welches für die Bearbeitung der Benutzeranfragen zuständig ist. In ihn können über einen Plugin-Mechanismus *Fuzzy-Operatoren* und *Fuzzy-Evaluatoren* integriert werden. *Fuzzy-Operatoren* sind verallgemeinerte Konzepte der bekannten booleschen Quantoren UND und ODER, die in Abschnitt 4.2 näher betrachtet werden sollen. Sie bilden einen zentralen Bestandteil des Systems und übernehmen die Berechnung der Rankings der Dokumente. Die *Fuzzy-Evaluatoren* hingegen sind Implementationen von Auswertungsstrategien. Ihnen kommt eine besondere Bedeutung zu, da die beim gewichteten Retrieval zu bearbeitenden Ergebnismengen in der Regel sehr umfangreich sind - häufig handelt es sich um mehrere zehn- bis hunderttausend Dokumente. Eine effiziente und optimierte Bearbeitung dieser Mengen ist daher unumgänglich. Je nach Komplexität und Struktur der Anfragen werden verschiedene Evaluatoren ausgewählt, die jeweils die optimale Auswertungsstrategie für die konkrete Anfrage implementieren. Darauf soll in Abschnitt 4.3 genauer eingegangen werden.
Die *Anbindung an die Datenbanken* geschieht über sog. Wrapper. Da der Rechercheassistenten als Meta-Suchmaschine konzipiert wurde, ist es wichtig, daß unterschiedliche, heterogene Datenbanken möglichst leicht in das System integriert werden können. Trotz der technisch teilweise sehr unterschiedlichen Zugriffsarten, Protokollen und Anfragesprachen muß der Fuzzy-Query-Interpreter über eine einheitliche Schnittstelle mit den Datenbanken kommunizieren können. Dies ist Aufgabe der Wrapper, die als „Übersetzer" dieser unterschiedlichen Zugriffsarten fungieren.

4.2 Fuzzy-Operatoren

Wie in Abschnitt 3.2 anhand eines Beispiels verdeutlicht wurde, verhalten sich die booleschen Quantoren UND und ODER zu restriktiv bzw. zu undifferenziert. Da die booleschen Quantoren keine Gewichte erlauben, kann das *Ranking* eines Dokumentes allein in Abhängigkeit des *Anteils* der enthaltenen Stichworte an den vom Benutzer eingegebenen Stichworte berechnet werden. Abbildung 6 stellt diesen funktionalen Zusammenhang graphisch dar: Die beiden booleschen Quantoren liefern lediglich die *diskreten* Werte 0 oder 1 entsprechend den Wahrheitswerten FALSE und TRUE. Bei dem ODER-Operator genügt bereits *ein* enthaltenes Stichwort, um dem Dokument eine volle Relevanz zuzuordnen, wohingegen bei UND *alle* Stichworte enthalten sein müssen, um ein Ranking größer 0 zu erhalten. Bei keinem der beiden Operatoren findet ein *graduelles* Ranking statt - Dokumente, die z.B. zwei Stichworte enthalten, bekommen dasselbe Ranking zugewiesen wie Dokumente, die 90% aller eingegebenen Stichworte enthalten.
Wünschenswert sind Quantoren, die entsprechend dem natürlichen Sprachgebrauch Konzepte wie ,,einige", ,,viele" oder ,,möglichst alle" realisieren. Dies wurde mit Fuzzy-Operatoren, wie sie - vereinfacht dargestellt - in Abbildung 7 zu sehen sind, erfolgreich realisiert. Es handelt sich dabei um Generalisierungen der booleschen Quantoren. Fuzzy-Operatoren erlauben ein *graduelles* Ranking, d.h. es werden einem Dokument - wieder in Abhängigkeit des Anteils der enthaltenen Stichworte - *kontinuierliche* Relevanzeinstufungen zwischen 0 und 1 zugeordnet. Der Operator ,,viele" z.B. verhält sich linear und ordnet einem Dokument, welches die Hälfte aller eingegebenen Stichworte enthält, eine mittlere Relevanz von 0.5 zu. Der Operator ,,möglichst alle" hingegen ist deutlich restriktiver und teilt einem solchen Dokument lediglich eine Relevanz von etwa 0.06 zu. Bei einer Anfrage mit diesem Operator würde dieses Dokument also relativ weit am Ende der Ergebnisliste präsentiert werden.
Durch die Möglichkeit, jedem Suchbegriff ein Gewicht zuzuordnen, wird die mathematische Definition eines Fuzzy-Quantors jedoch noch weitaus komplexer als es der natürliche Sprachgebrauch - in dem eine ,,Fuzziness" implizit immer vorhanden ist - nahelegt. Eine graphische Darstellung wie die der booleschen Operatoren in Abbildung 6 ist hier nicht mehr ohne weiteres möglich. Die Grafik in Abbildung 7 stellt lediglich eine Projektion der mehrdimensionalen Funktion auf $w_i = 1.0$ dar - also den Spezialfall, daß alle Gewichte auf 1.0 (entsprechend ,,sehr wichtig" im natürlichen Sprachgebrauch) gesetzt wurden. Durch die Möglichkeit, daß der Be-

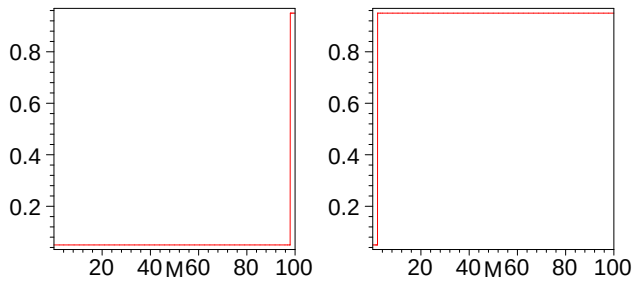

Abbildung 6: Die booleschen Aggregationsoperatoren ,,UND" und ,,ODER", Ranking eines Dokumentes in Abhängigkeit des Anteils der enthaltenen Stichworte. Das Ranking eines Dokuments bei einer Verknüpfung der Stichworte mit UND oder ODER nimmt lediglich die diskreten Werte 0 und 1 an, und erlaubt aufgrund fehlender Abstufung keine Sortierung.

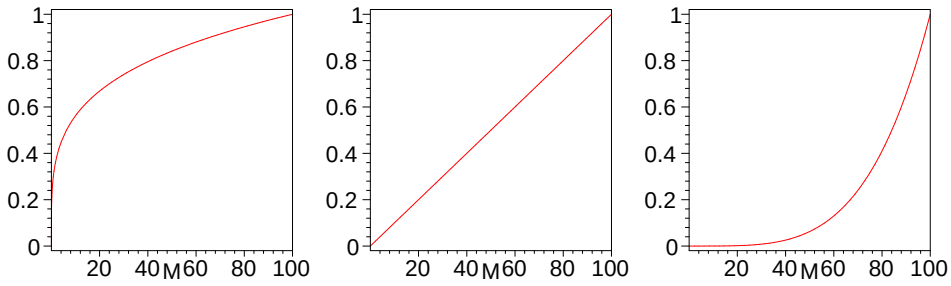

Abbildung 7: Vereinfachte Darstellung der Fuzzy-Aggregationsoperatoren ,,einige", ,,viele" und ,,möglichst alle", Ranking eines Dokumentes in Abhängigkeit des Anteils der enthaltenen Stichworte. Im Gegensatz zu den booleschen Aggregationsoperatoren ist hier das kontinuierliche Spektrum an Rankings zwischen 0 und 1 komplett abgedeckt, so daß eine wohldefinierte Sortierung der Dokumente nach Relevanz möglich ist.

nutzer den einzelnen Stichworten eine unterschiedliche Relevanz zuordnen kann (von 0.1 entsprechend ,,eher unwichtig" bis 1.0 entsprechend ,,sehr wichtig" - vgl. Abb. 1), wird erreicht, daß die Rankings der Dokumente noch exakter auf die Benutzeranfrage abgestimmt werden können: Sucht ein Benutzer z.B. nach Dokumenten über ,,information retrieval" und ,,fuzzy logic", und ist die erste Phrase für ihn ,,sehr wichtig", die zweite jedoch nur ,,mittel wichtig", so erscheinen die Dokumente, die lediglich die Phrase ,,information retrieval" enthalten noch vor denen, die ausschließlich ,,fuzzy logic" beinhalten.[2]

4.3 Fuzzy-Evaluatoren

Wie eingangs erläutert können die beim gewichteten Retrieval zu berechnenden Teilergebnismengen leicht einen sehr großen Umfang von bis zu mehreren Hunderttausend Dokumenten bekommen. Um diese Datenmengen effizient bearbeiten zu können, und kurze Antwortzeiten zu garantieren, wurde auf das sog. *Streaming*-Prinzip zurückgegriffen. Dieses setzt voraus, daß nicht alle Ergebnisse zu einem bestimmten Zeitpunkt auf einmal vorliegen müssen, sondern *iterativ* - also auf Anfrage hin nach und nach - Ergebnisse geliefert werden können. Die *Streams* müssen dabei sortiert sein, daß heißt, dem Benutzer werden zuerst die Dokumente mit dem höchsten Ranking präsentiert.
Je nach Struktur und Komplexität einer Benutzeranfrage gilt es, verschiedene Teilergebnis-Streams zu einem Gesamtergebnis-Stream zusammenzuführen. Durch die notwendige Aufrechterhaltung der Sortierung sind hier unterschiedliche Strategien nötig, um eine optimale Evaluation dieser Streams zu erzielen. Die *Fuzzy-Evaluatoren* sind Implementierungen dieser Strategien und werden je nach Art der Anfrage ausgewählt.
Ein Beispiel soll die Komplexität des Sachverhaltes verdeutlichen: Sucht ein Benutzer nach Dokumenten, die *,,möglichst viele"* der Suchbegriffe ,,information retrieval", ,,fuzzy logic" und ,,quantor" enthalten[3], so werden - wie im vorigen Abschnitt erläutert - nicht nur Dokumente nachgewiesen, die *alle* dieser Suchbegriffe enthalten, sondern z.B. auch Dokumente, die nur die Begriffe ,,information retrieval" und ,,fuzzy logic" enthalten, nicht aber den Begriff ,,quantor". Diese Dokumente erhalten

[2] Dokumente, die beide Suchphrasen enthalten, erscheinen selbstverständlich nach wie vor mit einem Ranking von 1.0 an erster Stelle.

[3] Der Einfachheit halber soll hier wieder auf die Verwendung von Gewichten verzichtet werden.

selbstverständlich ein niedrigeres Ranking, als Dokumente, die *alle* Begriffe enthalten. Spezifiziert der Benutzer jedoch in seiner Anfrage zudem, daß er besonders an *„möglichst neuen"*[4] Dokumenten interessiert ist, so erhält ein Dokument, welches sehr aktuell ist, aber nur zwei der drei eingegebenen Suchbegriffe enthält, u.U. ein höheres Ranking als ein Dokument, in dem zwar alle drei Suchbegriffe auftauchen, das aber schon sehr alt ist.
An diesem Beispiel ist zu erkennen, daß es nicht ausreicht, zunächst ausschließlich diejenigen Dokumente zu betrachten, die alle Suchbegriffe enthalten, genauso wenig wie es genügt, die Dokumente ausschließlich nach Erscheinungsjahr sortiert zu betrachten. Stattdessen ist eine holistischere Berücksichtigung *aller* Eigenschaften eines Dokumentes nötig, um effizient das Ergebnis berechnen zu können. Die Verknüpfung der verschiedenen Teilergebnis-Streams ist also keineswegs trivial.
Für die im Beispiel gezeigte, und für bestimmte andere Klassen von Anfragestrukturen konnten *Fuzzy-Evaluatoren* entwickelt werden, die nachgewiesenermaßen eine optimale Auswertungsstrategie repräsentieren.

4.4 Implementierung

Für die Implementierung des Rechercheassistenten wurde der plattformübergreifende CORBA-Standard verwendet. Dieser ermöglicht die Entwicklung objektorientierter Anwendungen, und vereinfacht die Integration von Teilkomponenten, die in unterschiedlichen Programmiersprachen implementiert werden. Die einzelnen Module des Gesamtsystems können durch die Verwendung von CORBA sehr einfach auf verschiedene Rechner verteilt werden, wodurch ein hohes Maß an Skalierbarkeit gewährleistet ist.
Das Grundsystem, der Fuzzy-Query-Interpreter, wurde zusammen mit den Wrappern aus Gründen der Performanz in C++ entwickelt. Es läuft auf vier Rechnern des Typs Athlon mit 600 MHz und einem Arbeitsspeicher von jeweils 128MB RAM, die mit dem Betriebssystem Linux (SuSe 7.2) arbeiten.
Das User-Frontend wurde aus eingangs geschilderten Gründen nicht als reines HTML-Formular entwickelt, sondern als Java-Applet realisiert. Durch die Beschränkung auf das Java Development Kit 1.02 wurde ein sehr hohes Maß an Portabilität sichergestellt, so daß das Interface auf allen gängigen Browser/Betriebssystem-Kombinationen lauffähig ist.
Als Bindeglied zwischen dem User-Frontend und dem Fuzzy-Query-Interpreter dient ein Java-Servlet, welches Server-seitig läuft und mit dem Applet ausschließlich über HTTP/HTML kommuniziert. Es stellt damit die Brücke zum Fuzzy-Query-Interpreter dar, mit dem es über die CORBA-Schnittstelle verbunden ist. Durch die Verwendung dieses Servlets konnte das Applet soweit abgespeckt werden, daß auch für Modem-Benutzer, bei denen der Netzwerkzugang den Engpaß darstellt, ein schneller Zugang gewährleistet wird.

5 Weiterer Projektverlauf

Entsprechend dem Arbeitsprogramm stehen noch folgende Arbeitspakete aus:

- Evaluation und Tests
- Integration weiterer Datenbanken
- Abschlußdokumentation

[4] Auch hierbei handelt es sich um einen Fuzzy-Operator, der ein graduelles Ranking - in diesem Fall in Abhängigkeit des Alters eines Dokumentes - vornimmt.

5.1 Evaluation und Tests

Der Rechercheassistent liegt als eine stabile, grundsätzlich evaluierbare und testfähige Anwendung vor, auf die bereits häufig zugegriffen wird (über 800 Aufrufe der Suchmaske in zwei Wochen vorlesungsfreier Zeit). Überlegungen zum zweckmäßigen Datenbankangebot finden sich in Abschnitt 5.2. Die Evaluation erfolgt im Rahmen des bewilligten Projektumfangs (der bewilligten Mitarbeiterstellen) gemäß Förderungsantrag. Drei Anmerkungen sollen hierzu gemacht werden:
Die Auswertung von Email-Feedback, zu dem die Benutzer ermuntert werden, soll verstärkt in die Evaluation einbezogen werden. Hierfür wurde ein besonderer Briefkasten eingerichtet und in die Frame-Struktur des Rechercheassistenten eingebettet.
Die Implementierung der Protokollfunktionen nach AP10 zur statistischen Analyse der Log-Dateien wurde zeitlich vorgezogen und ist bereits jetzt weitgehend abgeschlossen.
Zur Zeit ist noch unklar, ob und wie die der Bibliothek in AP9 zugedachten Aufgaben durch diese erledigt werden können, da die Präparation von Labortests keine genuine Bibliotheksaufgabe ist und die Bestimmung der ,,Recall Base" extrem aufwendig ist, da hier auch die *nicht* nachgewiesenen Dokumente auf Relevanz geprüft werden müssen. Es wäre sehr hilfreich, wenn man hierzu auf bereits existierende Datenbestände und Erhebungen zurückgreifen könnte.

5.2 Integration weiterer Datenbanken

Die Integration weiterer Datenbanken wird im Arbeitsprogramm des Antrags unter zwei Aspekten angesprochen: ,,AP11: Skalierbarkeitstest" mit Integration der BRS-Datenbanken JADE und Periodical Contents Index / Chadwyck-Healey und ,,AP12: Evaluation der Offenheit bzgl. Integration heterogener Datenbanken" mit Integration der ,,Fremddatenbanken" Springer LINK und ACADEMIC Press. Damit wird ein suchbares Volumen von insgesamt über 25 Mio. Dokumenten erreicht. Es ist aber nicht sinnvoll, die Integration dieser weiteren Datenbanken nur isoliert unter den genannten Aspekten zu betrachten: Benutzer, die den Rechercheassistenten nicht als Tester, sondern als Informationssuchende nutzen, werden sich nur in dem Maße zu den Möglichkeiten und ggf. Defiziten des Rechercheassistenten äußern (können), als über diesen Informationen zu ihren jeweiligen Interessengebieten in nennenswertem Umfang angeboten werden. Es sollte daher eine fächerübergreifende Attraktivität des Datenbankangebots gegeben sein. Auch die Ergebnisse des operationalen Tests (AP10) werden nur in dem Maße als typisch angesehen werden können, in dem es gelingt, ein auch inhaltlich motiviertes Interesse an der Nutzung des Rechercheassistenten zu erzeugen, das über das kurzzeitige ,,Ausprobieren" der Suchtechnik hinausgeht. Aus diesen Gründen wird in Kürze der Periodical Contents Index integriert werden und wird auch die Integration der weiteren Datenbanken zügig erfolgen, um eine möglichst lange Nutzungsphase mit mehr oder weniger vollständigem Datenbankangebot in die Evaluation bzw. Tests miteinbeziehen zu können.

5.3 Abschlußdokumentation

Auf Grundlage des Rechercheassistenten konnten bereits viele theoretische und praktische Kenntnisse zum gewichteten Retrieval, der Fuzzy-Logik, Anfrageoptimierung, zu effizienten Strategien der Anfrageabarbeitung, Problemstellungen und Aspekten des Meta-Suchmaschinen-Konzeptes, ergonomischer Benutzeroberflächengestaltung, modularisierter und verteilter, skalierbarer Anwendungen, server-seitiger Logging-Mechanismen zur Analyse des Benutzerverhaltens und anderer Problemstellungen gewonnen werden. Neben einer technischen Systembeschreibung des Re-

chercheassistenten sollen diese gemäß AP13 in einer Abschlußdokumentation zusammengefaßt werden.

6 Weitere Schritte und Ausblick

Die jetzt ins Netz gestellte Pilotanwendung ist eine von zwei Alternativen, die im Rahmen des Projekts entwickelt wurden: die öffentlich angebotene verfügt über die vollen Suchfunktionalitäten, die dem Benutzer die Eingabe auch komplexer Suchanfragen erlaubt. Die zweite - ebenfalls funktionsfähige - Suchmaske unterscheidet sich von ihr durch eine vereinfachte Suchmaske mit festen Voreinstellungen für den Aggregationsoperator und die Termgewichte, die den Benutzern verborgen sind. Der mögliche Einsatzbereich für diese vereinfachte Version ist die im Seitenangebot der UB Bielefeld angebotene Metasuche - de facto eine Suche in der Digitalen Bibliothek NRW, die aber nahtlos in das Seitenangebot der Bibliothek integriert ist. Ein alternativer Einsatzbereich wäre eine Einbindung in den regulären OPAC - dieser noch mehr als die ,,Digitale Bibliothek" ein System des Massennachweises. Daß gerade hier - bei stark den genutzten Systemen - der genuine Einsatzbereich der vereinfachten Suche liegt, bedarf keiner besonderen Begründung. Über den Zeitpunkt des Einsatzes im öffentlichen Angebot ist noch nicht entschieden. Daß der Version mit der vollen Funktionalität in der Pilotphase andererseits der Vorzug zu geben war, liegt auf der Hand, da Novität und Effizienz der neuen Suchtechnik so am wirkungsvollsten demonstriert (und auch evaluiert) werden können; schließlich erhält man auch mit Internet-Suchmaschinen ,,relevanzsortierte" Trefferlisten.
Die besonders einfache Bedienbarkeit der noch nicht eingesetzten vereinfachten Version - bei einer Massennutzung der dominierende Gesichtspunkt - wird (wegen der Voreinstellungen) durch Abstriche bei der optimalen Anfrageformulierung erkauft. Um so interessanter sind die Perspektiven, die sich in diesem Kontext aus der intelligenten Interpretation der Benutzeranfragen ergeben, die (u.a.) Gegenstand eines inzwischen eingereichten Folgeantrags sind. Mit der funktionierenden Plattform des Rechercheassistenten ist es möglich, durch relativ geringen Zusatzaufwand weitere Leistungssteigerungen durch verschiedene Ansätze zu realisieren. Der Bedienungskomfort und die Retrieval-Qualität sollen in Zukunft durch eine intelligente Interpretation und automatische Erweiterung der Benutzeranfrage noch weiter gesteigert werden. Neue Anfrage-Evaluations-Strategien werden dies unterstützen und neben einer Performance-Steigerung eine noch feinere Granulierung des Rankings und eine exaktere Relevanzeinstufung ermöglichen. Ein Profildienst mit integriertem Alert-Service wird die regelmäßige Recherche nach Neuerscheinungen zu Interessengebieten überflüssig machen - das System führt diese stattdessen automatisch durch und benachrichtigt den Benutzer per Email.
Wenn sich die Erwartungen in das Folgeprojekt erfüllen, wird es möglich sein, eine Suche mit einfachster - beinahe trivialer - Bedienbarkeit mit einer Vielzahl den Recherchegegenstand charakterisierenden Suchbegriffen und (fast) optimaler Sucheffizienz zu realisieren.

7 Unterschrift

Projektpartner:

(Dr. Neubauer, Ltd. Bibliotheksdirektor)	gez. Knoll (Prof. Dr. Knoll)