Osiris: Qualitative Fortschritte bei der Literaturrecherche

Marc Ronthaler
Marc.Ronthaler@CL-KI.uni-osnabrueck.de
Institut für Semantische Informationsverarbeitung
Universität Osnabrück
49069 Osnabrück

Vortrag im Rahmen der *Informatik ´98*,
28. Jahrestagung der Gesellschaft für Informatik
21.09.98-25.09.98 in Magdeburg

Veröffentlicht in: Dassow, J./ Kruse, R. (Hrsg.):
Informatik '98 : Informatik zwischen Bild und Sprache ;
28. Jahrestagung der Gesellschaft fuer Informatik; Magdeburg,
21. bis 25. September 1998. Berlin [u.a.] : Springer, Informatik aktuell, 1998

Osiris: Qualitative Fortschritte bei der Literaturrecherche

Marc Ronthaler
Marc.Ronthaler@CL-KI.uni-osnabrueck.de
Institut für Semantische Informationsverarbeitung
Universität Osnabrück
49069 Osnabrück

Zusammenfassung Nachdem in beinahe allen großen Bibliotheken der alte Zettelkatalog durch den OPAC[1] abgelöst wurde, sind die Nachteile dieses Systems inzwischen offenkundig. Die Suchmöglichkeiten sind nicht am Bedarf des Endbenutzers orientiert und die Suchergebnisse oft interpretationsbedürftig, womit Endbenutzer überfordert sind. Statt in Benutzerschulungen zu investieren, will das OSIRIS[2]-System die Literaturrecherche mit natürlicher Sprache auf einer automatisch generierten Wissensbasis ermöglichen. Die mit OSIRIS erzielten Ergebnisse sind zudem von besserer Qualität, die Zahl der relevanten Treffer liegt in der neuen Version 2.0 oft um den Faktor 10 über der des OPAC.

1 Einleitung

Seit Mitte 1996 fördert die Deutsche Forschungsgemeinschaft (DFG) mit dem Projekt OSIRIS an der Universitätsbibliothek Osnabrück die Neukonzeption des Bibliothekszugangs mit Hilfe multilingualer, natürlichsprachlicher Retrievaltechniken.
Nach einer kurzen Darstellung der besonderen Problematik des Anwendungsbereiches „Universitätsbibliothek“ wird es um die Neuerungen gehen, die nach sechs Monaten hochschulöffentlichen Betriebes in die nun vorliegende Version 2.0 eingegangen sind.
Im Anschluß daran soll anhand einiger Beispiele der Fortschritt demonstriert werden, den die in OSIRIS eingesetzten Techniken dem Bibliotheksbenutzer bringen.

2 Die Problemlage

In den vergangenen Jahren wurde in den meisten wissenschaftlichen Bibliotheken der klassische Zettelkatalog als Recherchemöglichkeit für den Endbenutzer vom sog. OPAC abgelöst. Ein typischer OPAC, und dazu zählt das Osnabrücker System, ist über eine Telnetverbindung mit VT100-Oberfläche zu erreichen. Der Benutzer kann nach folgenden Aspekten suchen: *Titelstichwörter*, *Personennamen*, *Körperschaften*, *Kongreßtitel*, *Serientitel*, *Systematik*, *Nummern* (ISBN etc.) und *Signaturen*.[3]

[1] **O**nline **P**ublic **A**ccess **C**atalog
[2] **O**snabrück **I**ntelligent **R**esearch **I**nformation **S**ystem
[3] Auf die besondere Problematik des Suchaspekts *Schlagwort*, der an einigen Bibliotheken angeboten wird, wird weiter unten eingegangen.

In vielen Bibliotheken ist der OPAC inzwischen die einzige Möglichkeit, im aktuellen Bestand zu recherchieren. So wird in der Universitätsbibliothek Osnabrück, mit ca. 1 Million Büchern eine mittelgroße Bibliothek, der Zettelkatalog seit 1996 nicht mehr aktualisiert. Laut einer Nutzerbefragung, die im Rahmen der die OSIRIS-Entwicklung begleitenden Evaluationsmaßnahmen[4] durchgeführt wurde, wird der Zettelkatalog von vielen immer noch dem OPAC vorgezogen, auch wenn die gesuchten Titel bereits im OPAC verzeichnet sind. Ein Grund hierfür ist in der ausgesprochen benutzerunfreundlichen VT100-Oberfläche zu sehen. Deshalb gibt es zahlreiche Ansätze, die OPAC-Oberflächen zu überarbeiten, z.B. sie WWW-konform zu gestalten. Doch auch eine sinnvolle, an softwareergonomischen Gesichtspunkten orientierte Oberflächengestaltung ändert nichts an der eigentlichen Schwäche des OPAC-Systems als Rechercheinstrument für den Endbenutzer: Der OPAC ist auf die Bedürfnisse der Bibliothekare zugeschnitten. Endbenutzer, die den OPAC zur Literaturrecherche benutzen, verstehen nachweislich die Bedeutung der meisten Suchmöglichkeiten nicht.[5] In der Folge verwenden sie daher beinahe ausschließlich nur die Suchaspekte *Titelstichwörter* und *Personennamen*.

Es ist zwar einfach, mit diesen beiden Optionen das Vorhandensein eines Werkes, dessen Titel und Verfasser sowie deren genaue Schreibweise bekannt sind, zu überprüfen, die Suche nach Literatur zu einem Thema aber gestaltet sich damit frustrierend. Die Ergebnisse solcher Suchanfragen sind sehr oft viel zu groß, oft so groß, daß das OPAC-System deren Darstellung verweigert (Bsp. Stichwort *Chemie*, *Geschichte* oder *Deutschland* an einem deutschen OPAC).

Ebenso leicht geschieht es, daß Benutzer überhaupt kein Ergebnis erhalten. Da Stichwortsuche im OPAC reines pattern matching ist, haben nur die Suchanfragen Erfolg, die genau die im Titel vertretenen Ausdrücke verwenden. Darüber hinaus befindet sich kaum weiteres sprachliches Material in den Datensätzen, so daß laut Schulz[6] jede zweite bis dritte Stichwortsuche überhaupt keinen Treffer erzielt. Hinzu kommt die mangelnde Robustheit und Flexibilität gegenüber den Benutzereingaben: Allein aufgrund unerkannter (!) Fehler in der Eingabe enden nach Schulz 10% aller Anfragen mit einem Nulltrefferergebnis.

Benutzer, die mit Hilfe von Titelstichwörtern Literatur zu einem Thema suchen und denen es gelingt, durch geschickte Formulierung sowohl leere als auch übergroße Ergebnismengen zu vermeiden, geraten in die eigentlich bedrohlichste Lage: Wie möchten sie sichergehen, daß die gefundenen Titel wirklich alle für die Themenstellung relevanten Titel im Bestand sind?

Sucht ein Benutzer im OPAC der Bibliothek Osnabrück bspw. nach Literatur zum Thema *Sport im Alter*, so ist er gezwungen, einen Ausdruck in der Booleschen Logik der Anfragesprache zu konstruieren, etwa wie folgt: `sport AND alter`. Das Ergebnis dieser Anfrage sind fünf Bücher. Benutzer, die erwartungsgemäß dem System vertrauen, ärgern sich, daß ihre Interessen nicht in der Beschaffungspolitik der Bibliothek repräsentiert sind. Benutzer, die (zu Recht)

[4] Siehe auch Abschnitt 5.
[5] vgl. [Dreis(1994)] und [Schulz(1994)]
[6] [Schulz(1994), :299]

an der Zuverlässigkeit des Systems zweifeln, beginnen nun, Ausweichstrategien zu entwickeln, um die Bücher zu finden, die der OPAC ihnen vorenthält. Eine Suche im semantischen Umfeld mit *Seniorensport* oder *Alterssport* fördert schon mehr zu Tage, aber auch mit den geschicktesten Strategien werden sie nie sicher sein können, alle für ihre Suchanfrage inhaltlich relevanten Werke der Bibliothek gefunden zu haben.

2.1 Traditionelle Lösungsansätze

Angesichts der genannten Probleme wurden bislang zwei Lösungsstrategien angeführt, *Schlagwörter* und *Trunkierung*. Beide Vorschläge sind allerdings untauglich, die beschriebene Lage zu verbessern.
Einige Bibliotheken bieten ihren Benutzern im OPAC die Möglichkeit, mit einer Suchoption *Schlagwort* nach inhaltlichen Kriterien und nicht nach Stichwörtern zu suchen. Schlagwörter werden intellektuell durch ausgebildete Fachkräfte einem Datensatz angefügt. Die hohen Kosten dieser personalintensiven Erschließung haben dazu geführt, daß nur etwa 50% der Bestände der Universitätsbibliothek Osnabrück unsystematisch mit Schlagwörtern versehen sind. In dieser Hinsicht unterscheidet sich die Universitätsbibliothek Osnabrück nicht von den meisten anderen deutschen Bibliotheken. Eine Suchoption *Schlagwort* arbeitet also von vornherein nur auf einem Bruchteil des eigentlichen Bestandes, weshalb in Osnabrück auf eine solchen Option im OPAC verzichtet wurde. Auch erscheint eine Benutzung der Schlagwörter aufgrund der enormen Komplexität der Ansetzungsregeln[7] durch einen bibliothekarischen Laien kaum möglich. Nur eine regelkonforme Anwendung garantiert aber Erfolg: *Hund* trifft keine inhaltlich relevanten Einträge, denn die Regel vom „engen Schlagwort“ sorgt dafür, daß die entsprechenden Datensätze mit *Dackel*, *Terrier* usw. versehen werden. Da aber auch auf Schlagwortmaterial nur mit pattern matching gesucht wird, bleibt auch in einem zu hundert Prozent verschlagworteten Bestand die Suche nach Literatur zu *Hund* ohne Erfolg.
Trunkierung, das ist die Suche mit sog. Wildcards oder Jokerzeichen, um beliebige Wortendungen zu treffen, kann von den Benutzern nicht erwartet werden, denn diese haben „erhebliche Schwierigkeiten mit der Verwendung von Maskierungsmöglichkeiten“[8] und würden sie daher meiden. Schulz schreibt: *„Der Zweck von Trunkierungsmöglichkeiten wird offensichtlich nicht verstanden, die Funktion folglich nicht genutzt.“*[9]
Außerdem ändert Trunkierung nichts am Fehlen von Schlagwörtern und führt in der Stichwortsuche noch eher zu den erwähnten übergroßen Treffermengen. Die Ergebnisse werden durch Trunkierung darüber hinaus sehr unpräzise, weil *beliebige* Wortfortsetzungen getroffen werden. Für semantische Phänomene (bspw. Synonyme und Ambiguitäten) stellt Trunkierung ebenfalls keine Lösung dar.
In dieser Situation, die durch bloße Oberflächengestaltung nicht zu verbessern ist, erscheint es angesichts der Kosten einer möglichen Vervollständigung der

[7] **R**egeln für den **S**chlag**w**ort**k**atalog
[8] [Schulz(1994), :300]
[9] siehe dort

sachlichen Erschließung sowie derem zweifelhaften Nutzen für die Recherche sinnvoll zu versuchen, auf der Grundlage des vorhandenen Datenbestandes eine qualitative Verbesserung der Sachrecherche des Endbenutzers zu erreichen. Dies ist der Ausgangspunkt des OSIRIS-Systems.

3 Das Osiris-System

OSIRIS trägt mit zwei Dingen zur Verbesserung der geschilderten Lage bei: Einerseits mit einer robusten, natürlichsprachlichen Endbenutzerschnittstelle zur Datenbank und andererseits mit einer intelligenten, automatischen Aufbereitung des verfügbaren Datenbestandes. In dieser Kombination und mit der Betonung automatischer, regelbasierter Verfahren, hebt sich das OSIRIS-System von allen uns bekannten Ansätzen zur Rechercheverbesserung ab.

3.1 Die Eingabeverarbeitung

Die in OSIRIS eingesetzte Endbenutzerschnittstelle hat die Aufgabe, natürlichsprachliche Anfragen des Benutzers syntaktisch und semantisch so zu analysieren, daß mit dem Ergebnis der Analyse eine Suche auf der Datenbank erfolgreich durchgeführt werden kann. Phrasen müssen erkannt und die sie konstituierenden Teile semantisch zueinander in Relation gesetzt werden. Komposita müssen in ihre Bestandteile zerlegt, flektierte Formen müssen normalisiert werden. Je nach Ansetzungspraxis muß Plural auf Singular (für das Deutsche) oder Singular auf Plural zurückgeführt werden. Weiterhin müssen, als Vorbereitung einer eventuell mehrstufigen Datenbankanfrage, Synonyme, Hypo- und Hyperonyme sowie eventuell einfache Kategorieninformationen bereitgestellt werden.
Die verwendeten Komponenten sind auf ihr Einsatzgebiet in der Bibliothek hin ausgelegt: Morphologie, Grammatik und Semantik wenden einfache und effiziente Techniken an und nutzen die Vorteile des prinzipiell eingeschränkten Aufgabenbereiches.
Um robust gegenüber fehlerhaftem Input zu sein, muß OSIRIS die häufigsten und einfachsten Eingabefehler erkennen und selbsttätig beheben können. Fehler dieser Art sind einfache Buchstabenvertauschungen, Verdoppelung einzelner Buchstaben sowie Fehlschreibungen durch Benutzung von auf der Tastatur benachbarten Tasten. Insbesondere wird OSIRIS so tolerant gegenüber beliebigen Reformen der Rechtschreibung. Ohne Änderung der Daten ist OSIRIS in der Lage, beispielsweise *Schifffahrt* (mit drei *f*, geplante neue Rechtschreibung) auf *Schiffahrt* (bislang mit zwei *f*) abzubilden.
Flexibilität meint aber auch, daß OSIRIS dem Benutzer im Bereich der notorisch schwierigen Schreibung von Namen hilft. Ganz im Gegensatz zu klassischen OPAC-Systemen ist es in OSIRIS möglich, erfolgreich eine Autorin Namens *Isabelle Ebers* zu suchen, obwohl die betreffende Person *Isabella Schneider-Eberz* heißt. Darüber hinaus sind Benutzer nicht mehr gezwungen, wie OPAC-Systeme es vorschreiben, nach *Mulen, Alice van* zu suchen, sondern können Namen auch in der natürlichen Reihenfolge *Alice van Mulen* eingeben. Und dabei werden

sie auch noch Erfolg haben, wenn die gesuchte Person in Wirklichkeit *Alice ter Meulen* heißt.
Hinsichtlich des Sprachumfangs der zu erwartenden Eingaben stellt das Einsatzgebiet „wissenschaftliche Bibliothek“ in struktureller Hinsicht eine echte Einschränkung dar, in lexikalischer aber nicht. Die an OSIRIS gerichteten Anfragen werden stets Suchanfragen nach bestimmten Büchern oder Büchern zu einem Thema sein – OSIRIS ist kein allgemeines Bibliotheksauskunftssystem. Als Input sind also durchweg Nominalphrasen zu erwarten, wie sie als Titel wissenschaftlicher Arbeiten oft verwendet werden: *Mathematikunterricht in der gymnasialen Oberstufe*, *Subventionspolitik und Steuerreform* oder *Automatische Datenerfassung in deutschen Bibliotheken.* OSIRIS kann also problemlos eine eingeschränkte Syntax in der Anfrage voraussetzen, was auf der Oberfläche durch Vorgabe eines Satzanfanges realisiert wird. Im Fall der thematischen Literatursuche wird bspw. vom Benutzer erwartet, daß er den Satz *Ich suche Literatur zum Thema ...* vervollständigt. So wird intuitiv die Eingabe von Nominalphrasen angeregt, ohne daß der Benutzer in seinen Möglichkeiten beschränkt würde.
Trotz Ausbildung von Bestandsschwerpunkten muß in einer wissenschaftlichen Bibliothek aber mit Literatur zu *allen* Wissensgebieten und das in einem Spektrum vom einführenden Werk bis hin zu Tagungsbänden und Dissertationen gerechnet werden. Deshalb muß OSIRIS mit dem gesamten aktuellen wissenschaftlichen Vokabular vertraut sein. Die Schwierigkeiten an dieser Stelle liegen dabei nicht nur in der bloßen Menge des Vokabulars, sondern auch in der zu gewährleistenden Aktualität. Manuell zu wartende Wörterbücher erscheinen uns hier unrealistisch. Wir setzen auf regelbasierte Ansätze, inkrementelle und automatische Lexikonerweiterung, sowie in vielen Bereichen (wie z.B. der Kompositazerlegung) auf Heuristiken. So sind wir in der Lage, auch mit Lexika, die mehrere hunderttausend Einträge umfassen, noch effizient zu arbeiten.

3.2 Die Aufbereitung der Daten

Der zweite Aspekt, mit dem OSIRIS zur Verbesserung der Literaturrecherche beiträgt, ist die Generierung einer Wissensbasis, auf der im laufenden Betrieb gesucht wird. Diese Wissensbasis wird einmal automatisch generiert und später regelmäßig inkrementell erweitert. Die Informationen, die in dieser Wissensbasis zur Verfügung stehen, stammen im wesentlichen aus dem OPAC, in dem mehr Informationen enthalten sind, als sie der Endbenutzer normalerweise sieht. Neben den üblichen Angaben zu Verfasser und Titel finden sich im Idealfall noch deutsche wie fremdsprachige Schlagwörter, Notationen nicht lokaler Klassifikationssysteme und anderes mehr. Daß diese verborgen bleiben, liegt einerseits daran, daß der Endbenutzer bereits mit den jetzt im OPAC verfügbaren Möglichkeiten überfordert ist. Andererseits, und dies ist ein noch viel wichtigerer Grund, liegen die genannten Informationen zumeist nicht flächendeckend vor, wie am Beispiel der Schlagwörter bereits erörtert. Diese Informationen werden von OSIRIS vollständig gesammelt und ausgewertet.
Die Auswertung der Daten beruht darauf, daß bspw. ein Buch mit dem Titel *Lie groups and quantum mechanics*, das von einem Autor namens *Peter Cunning-*

ham veröffentlicht wurde, von einem Fachreferenten der Bibliothek in die lokale Systematik eingeordnet wird, in diesem Fall unter der Notation *TEN*. Dieselbe Arbeit aber hat in einer anderen Bibliothek eventuell auch jemand getan: So ist es gut möglich, daß z.B. in der Deutschen Bibliothek in Frankfurt dem Datensatz die Information angehängt wurde, daß das Buch inhaltlich mit den Themen *Quantenmechanik* und *Lie´sche Gruppe* zu tun hat. Eventuell wurde das Buch auch schon bei Erscheinen klassifiziert. Dann wird in der Library of Congress ein Mitarbeiter festgestellt haben, daß der Autor sich in diesem Werk mit den Gebieten *quantum mechanics* und *lie groups* beschäftigt.
OSIRIS wertet diese Informationen systematisch aus und stellt fest, daß von den Büchern, die in Osnabrück in der Systemstelle *TEN* stehen und Schlagwörter tragen, signifikant viele mit dem Schlagwort *Quantenmechanik* versehen sind. Deutlich weniger Exemplare tragen fremdsprachige Schlagwörter, unter diesen aber sind wiederum signifikant viele mit dem Schlagwort *quantum mechanics* versehen. Offenbar werden in der Systemstelle *TEN* die Bücher klassifiziert, die sich inhaltlich mit dem beschäftigen, was im Deutschen *Quantenmechanik* und im Englischen *quantum mechanics* genannt wird. Das gilt dann natürlich auch für die Bücher, für die keine Schlagwörter vergeben wurden, die aber ebenfalls unter *TEN* klassifiziert wurden. Damit OSIRIS auf dem gesamten Buchbestand suchen kann, ist es also nicht notwendig, daß alle Bücher tatsächlich mit Schlagwörtern versehen sind. Wir schätzen, daß ein Anteil von etwa 30% verschlagworteten Titeln ausreicht, damit OSIRIS erfolgreich eine Wissensbasis aufbauen kann.
Hilfestellung leisten bei den genannten Auswertungen natürlich die computerlinguistischen Komponenten, die für jede Suchanfrage neben der morphologischen Reduktion auch passende Synonyme, Hyponyme und Hyperonyme bereitstellen. So werden auch die genannten Nachteile einer schlagwortbasierten Suche umgangen. Diese computerlinguistische Funktionalität ist herkömmlichen bibliothekarischen Bemühungen um normgerechte Ansetzungen, normiertes Vokabular und manuelle Vervollständigung von Schlagwortmaterial hinsichtlich der Kosten, der Flexibilität, der Verfügbarkeit und Aufgabenangemessenheit weit überlegen.

3.3 Realisierung

OSIRIS ist modular aufgebaut: Alle Komponenten kommunizieren über Internet-Sockets miteinander, so daß die gesamte Anwendung verteilt in einem Netz unter Einbeziehung verschiedener Betriebssysteme (derzeit Linux und Solaris) arbeitet.
Seit der Version 2.0 arbeitet OSIRIS in der Eingabeverarbeitung auf einer Blackboardarchitektur, deren Kernkomponenten der Task Manager, der Parser und das Lexikon sind.
Der Task Manager ist für die syntaktische Vorverarbeitung verantwortlich. Er erhält die Benutzereingaben und kommuniziert mit dem Lexikon- und dem Parserserver. Zu den Aufgaben des Task Managers gehört dabei die wortweise Befragung des Lexikons sowie die Reduktion und Wiedervorlage von als unbekannt zurückgewiesenen Worten. Allgemein finden im Task Manager die Manipulationen statt, die nicht direkt während des Lexikonzugriffs realisiert werden können.

Die Ergebnisse des Lexikonprozesses werden zudem durch den Task Manager sortiert und bewertet und eine Auswahl dem Parserserver übergeben. Das Ergebnis des Parsingprozesses wird mit den Ergebnissen der lexikalischen Analyse vom Task Manager der OSIRIS-Wissensbasis übergeben.

Das Lexikon enthält Stammformeneinträge, wobei Endungen und ggf. erforderliche Umlautungen Bestandteil der lexikalischen Information im Eintrag sind. Beim Lexikonzugriff wird eine morphologische Reduktion vorgenommen, das Ergebnis ist die Zerlegung in Endung und Lemma, erkannte Affixe sowie Kongruenzinformationen. Kompositazerlegungen, Endungsanalyse und Korrektur einfacher Fehlschreibungen erfolgen ebenfalls in diesem Schritt. OSIRIS benötigt große Lexika mit mehr als 300.000 Stammformeneinträgen. Um diesen Anforderungen zu genügen, werden in OSIRIS die Lemmata als Suchschlüssel in einem Buchstabenbaum kodiert und die Inhalte der Einträge separat in einer Datenbank gespeichert. Dabei wird die morphologische Information von einer zugekauften Software erzeugt.[10]

Im Deutschen ist die Beziehung zwischen Konstituenten von Komposita sehr vage und kann daher nur auf der Basis von Heuristiken analysiert werden. So stützt sich die in OSIRIS verwendete Strategie auf die Beobachtung, daß der größtmögliche im Lexikon vorhandene rechte Rand eines Wortes als das spezifischste darin enthaltene bekannte Konzept das geeignete Objekt für die Suche in der Wissensbasis ist. Die Art der Beziehung zu den weiter links stehenden Teilen wird, falls bekannt, als eine unspezifische Modifikationsrelation interpretiert.

Zur Behandlung der notorisch schwierigen Semantik von Präpositionen wird ein einfaches Modell verwendet. Wenn Präpositionen mit einem zeitlichen oder räumlichen Aspekt annotiert sind, „restringieren“ sie das vom Nomen vertretene zentrale Konzept in Raum oder Zeit, andernfalls „modifizieren“ sie es. Anfragen wie „Datenverarbeitung in Bibliotheken“ werden daher semantisch genauso interpretiert wie „Datenverarbeitung in der Bibliothek“, nämlich als „Datenverarbeitung modifiziert-durch Bibliothek“.

Der Parserprozeß analysiert auf der Grundlage einer kontextfreien Grammatik mit Merkmalsannotationen und unter Berücksichtigung der durch den Task Manager gefilterten und sortierten Ergebnisse des Lexikonprozesses die syntaktische Struktur der Eingabe. Die Anforderungen, die OSIRIS an den Parser stellt, bestehen vor allen Dingen in einer optimierten Behandlung von Nominalphrasen sowie geringen Ressourcenansprüchen. Prinzipiell ist OSIRIS nicht auf die Verwendung eines bestimmten Parsers festgelegt. Derzeit wird eine Reimplementation des GEPARD-Parsers[11] eingesetzt, der bereits auf PCs auch mit umfangreichen Lexika und Grammatiken eine hohe Performanz erzielt. Dazu wird eine deklarative Grammatik mittels eines mehrstufigen parametrisierbaren Compilers in ein C-Programm übersetzt, so daß im Ergebnis ein sehr effizienter, für die jeweils gegebene Eingabegrammatik optimierter Chartparser zur Verfügung steht.

---

[10] „GERTWOL“ (deutsch) und „ENGTWOL“ (englisch) der Firma Lingsoft Inc., Helsinki, Finnland.

[11] [Langer(1996)]

In der OSIRIS-Wissensbasis werden mit den Ergebnissen der Eingabeanalyse passende Klassen und Titel gesucht. Das Ergebnis der Suche wird in Abhängigkeit von der Größe der Ergebnismenge bewertet und ggfs. die Darstellung einzelner Klassen unterdrückt. Entspricht die Eingabe einem Buchtitel, so wird dieser Titel auch dann präsentiert, wenn die entsprechende Klasse insgesamt nicht als zur Anfrage passend betrachtet wird. Personennamen sind phonetisch kodiert und in Namensbestandteile zerlegt abgelegt. Der Zugriff auf Namen ist somit weder an die lineare Ordnung der Eingabe noch (in gewissen Grenzen) an die korrekte Schreibweise gebunden. Auch ist es möglich, mit Teilen des Namens, z.B. bei Doppelnamen, erfolgreich zu suchen.
Die OSIRIS-Wissensbasis ist mit jeder Standard-SQL-Datenbank zu realisieren, für den Einsatz von OSIRIS in der Universitätsbibliothek Osnabrück verwenden wir z.Z. Oracle.

4 Beispiele für Retrievalverbesserung

Die deutlichen Verbesserungen der Recherchemöglichkeiten durch OSIRIS zeigen sich nicht nur in einem höheren Recall, sondern auch in der Präzision der Suchergebnisse. Im OPAC sind aufgrund von pattern matching Suchergebnisse hinsichtlich der Benutzeranfrage nicht immer inhaltlich relevant (siehe das Beispiel *Rom* weiter unten). Hingegen werden in OSIRIS Suchergebnisse in der Regel in Form von Klassen präsentiert, denen Bibliothekare Bücher aufgrund ihres Inhaltes zuordnen. Die sich daraus ergebenden Vorteile sollen anhand einiger Beispiele veranschaulicht werden.
So liefert OSIRIS für das bereits erwähnte Beispiel *Sport im Alter* die Klasse *Seniorensport* mit 49 Titeln, sowie die Klassen *Sport und Heilpädagogik* und *Motorisches Lernen*. Im OPAC trifft die Frage nach *Sport im Alter* gerade 5 Titel.
Das ebenfalls bereits angeführte Beispiel *Datenverarbeitung in Bibliotheken* führt im OPAC auf 5 Titel, OSIRIS präsentiert verschiedene Unterklassen von *Buch- und Bibliothekswesen*, darunter auch *Automatisierte Datenerfassung und Datenverarbeitung in der Bibliothek*. Ein Klick auf diese Klasse liefert bereits 38 Titel, die aufgrund der Klassenzugehörigkeit natürlich alle relevant sind. Ein weiterer Klick auf einen der Titel präsentiert denselben in der Vollanzeige. Hier besteht auch die Möglichkeit, das entsprechende Buch in eine eigene Literaturliste zu legen sowie zu überprüfen, ob es eventuell ausgeliehen ist.
Eine Anfrage wie *Chemie* wird am OPAC nicht beantwortet: Die Zahl der Treffer liegt im Bereich von mehreren Tausend, und solche Treffersets werden nicht mehr dargestellt. OSIRIS liefert für die Frage nach *Chemie* genau eine Antwort: Ein Symbol für die Klasse *Chemie*. Klickt der Benutzer auf das die Klasse repräsentierende Symbol, so findet er natürlich nicht alle Titel der chemischen Literatur vor, sondern die Unterklassen des Faches Chemie, *Organische Chemie*, *Anorganische Chemie* usw.
Sucht ein Benutzer nach *machine learning with neural networks*, so findet der OPAC nur ein Buch, in dessen Titel genau die gesuchten Stichworte vorkommen:

Machine learning: neural networks, genetic algorithms, and fuzzy systems. Alle anderen Bücher, insbesondere die deutschsprachigen, werden nicht gefunden. Das ist in der Praxis durchaus ein Problem, denn die deutschen Entsprechungen zu Fachtermini wie bspw. *loop space* oder *string theory* sind ungebräuchlich und deshalb dem Suchenden meist unbekannt. OSIRIS liefert auf die oben genannte Anfrage neben besagtem Werk, dessen Titel mit der Suchanfrage übereinstimmt, weitere Klassen aus Informatik, Künstlicher Intelligenz und Wirtschaftwissenschaften (Statistik). Diese Anfrage wird in OSIRIS nicht mit Hilfe einer Übersetzung bearbeitet, sondern mit den englischsprachigen Schlagwörtern, die in der OSIRIS-Wissensbasis vorhanden sind.

Sucht ein Benutzer im OPAC nach *Morphologie*, so findet er in Osnabrück 289 Titel. Vielen Benutzern ist zunächst unklar, daß ihre Suchbegriffe in anderen Disziplinen mit ganz anderer Bedeutung verwendet werden können (*Morphologie*, *Gruppentheorie* ...). Anhand der unstrukturierten Titelliste, die der OPAC als Ergebnis präsentiert, ist es für den Benutzer schwer zu erkennen, aus welchen Bereichen (Medizin, Biologie, Linguistik usw.) die Treffer stammen. Versucht er dann die hinsichtlich seiner Interessen inhaltlich relevante Teilmenge zu identifizieren, muß er jeden einzelnen Titel betrachten. Dieselbe Anfrage an OSIRIS liefert als Ergebnis die entsprechenden einschlägigen Unterklassen, u.a. für die *Computerlinguistik* die Klasse *Lexikalisches Wissen, Morphologie, Phonetik*, für die *Biologie* in der *Zoologie* und in der *Botanik* die entsprechenden Klassen, für die Einzelsprachen die jeweiligen Unterklassen *Morphologie* usw. usf.

Sucht ein Bibliotheksbenutzer nach Literatur zur Programmiersprache *C*, so wird er im OPAC nicht fündig: *C* ist ein sog. Stoppwort und wird von vornherein nicht zur Suche zugelassen. Die Suche in OSIRIS führt u.a. in der *Informatik* in die Unterklasse *Problemorientierte Programmiersprachen.*

Eine Suche nach Literatur zur Stadt *Rom* endet im OPAC mit einer Liste von 470 Titeln. Da die Suche auf pattern matching basiert, haben etwa 20% der gefundenen Titel nichts mit *Rom* zu tun: Ihnen liegt eine *CD-**Rom*** bei oder der Begriff wird im Titel genannt. OSIRIS liefert eine Klasse *Rom* als Ergebnis.

Neben der Themensuche und der oben bereits erwähnten Autorensuche, können in OSIRIS natürlich auch bestimmte Werke gesucht werden. Im Gegensatz zum OPAC reicht es hier allerdings, wenn man seine Anfrage wie in einer Buchhandlung notiert: *Shoham: Reasoning about Change* findet das entsprechende Werk, *Chomsky Syntactic Structures* findet neben dem englischen Original auch die deutsche Übersetzung.

5 Osiris in der Praxis

Die Laufzeit des OSIRIS-Projektes endet im August 1998. Seit dem Herbst 1997 steht das OSIRIS-System allen Mitgliedern der Universität Osnabrück zur Literaturrecherche zur Verfügung. Seit April 1998 ist die Version 2.0 universitätsweit freigegeben. Interessierte von außerhalb können sich unter `http://www.ub.uni-osnabrueck.de` registrieren lassen, um einen Zugang zu erhalten.

Neben der Literaturrecherche für Endbenutzer wird OSIRIS auch bei der Katalogisierung eingesetzt werden. Für neuerworbene Titel macht das System dem Bibliothekar einen Vorschlag, wie das betreffende Buch lokal klassifiziert werden soll. Die Informationen in der OSIRIS-Wissensbasis über den vorhandenen Bestand sowie eventuell vorliegende Fremddaten für das neu erworbene Buch, machen einen solchen Vorschlag möglich. Diese Technik wird bereits jetzt im experimentellen Betrieb eingesetzt: OSIRIS erzeugt automatisch eine nach Fachgebieten klassifizierte Liste der bestellten und der neu erworbenen Titel, die im WWW für alle Benutzer einsehbar ist.
Die DFG hat bereits Anfang des Jahres eine Verlängerung des Projektes bis Mitte 1999 beschlossen.

Seit Beginn des öffentlichen Betriebes wird das OSIRIS-System von Angehörigen des Fachbereiches Psychologie der Universität Osnabrück unter der Leitung von Dr. Kai-Christoph Hamborg nach software-ergonomischen Gesichtspunkten und hinsichtlich der ISO 9241/10 evaluiert. Ziel dieser Evaluation ist die konkrete Verbesserung der Oberfläche und der derzeitigen Funktionalität auf der Grundlage von Untersuchungen, die Aufschluß geben sollen über die Wünsche und Arbeitsstrategien der Benutzer. Dazu finden Befragungen von Endbenutzern wie von Experten statt, die die Recherchemöglichkeiten von OSIRIS sowie die des OPAC kommentieren sollen. Parallel dazu geht es in Laborexperimenten mit Versuchspersonen darum, anhand komplexer Aufgaben, Suchstrategien bei der Literaturrecherche experimentell zu ermitteln.
Darüber hinaus haben die Evaluationen das Ziel, im OSIRIS-Projekt einen langfristigen Entwicklungszyklus zu induzieren, der die Beobachtung des Einflusses von begleitenden Evaluationsmaßnahmen auf die Softwareentwicklung ermöglicht.

Danksagung Dieser Artikel beschreibt die Arbeit vieler Menschen. Ich danke Helmar Gust, Ingrid Recker, Claus Rollinger, Andreas Rother, Ulrich Sauer, Wilfried Teiken und Hartmut Zillmann für die Zusammenarbeit und hilfreiche Kritik. Für hilfreiche Kommentare danke ich weiterhin zwei anonymen Gutachtern.

Literatur

1. DREIS, G.: *Benutzerverhalten an einem Online-Publikumskatalog für wissenschaftliche Bibliotheken: Ergebnisse und Erfahrungen aus dem OPAC-Projekt der Universitätsbibliothek Düsseldorf.* 57 *Zeitschrift für Bibliothekswesen und Bibliographie, Sonderhefte.* Klostermann, 1994.
2. LANGER, H.: *Disambiguierung von Präpositionalkonstruktionen mit einem Syntaktischen Parser.* MEHL, S., A. MERTENS M. SCHULZ (): *Präpositionalsemantik und PP-Anbindung.* Uni Duisburg, 1996.
3. RECKER, I., M. RONTHALER H. ZILLMANN: *OSIRIS (Osnabrück Intelligent Research Information System) - ein Hyperbase Front End System für OPACs.* Bibliotheksdienst, 30(5):833–848, 5 1996.
4. SCHULZ, U.: *Was wir über OPAC-Nutzer wissen: Fehlertolerante Suchprozesse in OPACs.* ABI-Technik, 14(4):299–310, 1994.