Hartmut Zillmann
OSIRIS : Osnabrück Intelligent Research Information System

http://osiris.ub.uni-osnabrueck.de

Das OSIRIS-Projekt ist eine gemeinsames Projekt der *UB Osnabrück* und des *Instituts für Semantische Informationsverarbeitung* der Universität Osnabrück, das von 1996-1999 von der *Deutschen Forschungsgemeinschaft* im Rahmen ihres Bibliotheksförderprogrammes gefördert wird.

OSIRIS ist ein multilinguales intuitiv-natürlichsprachlich zu benutzendes Retrievalsystem, das die Anwendung klassischer Retrievaltechniken (spezielle Kommandosprache, Boolesche Verknüpfungen, Trunkierung etc.) für den Benutzer überflüssig macht und qualitativ bessere Suchergebnisse erzielt.

Bislang steht die in Datenbanken abgelegte, für Wissenschaft, Wirtschaft, Politik und Gesellschaft relevante elektronische Information praktisch nur dem Spezialisten klassischer Retrievaltechniken zur Verfügung. Ohne detaillierte Kenntnisse dieser Retrievaltechniken, wie über spezielle Kommandos, Boolesche Verknüpfungen, Trunkierung etc., ist eine effektive Recherche in weltweit verteilten Datenbanken für Nicht-Datenbankspezialisten unmöglich. Die Präsentation der ermittelten Suchergebnisse gestaltet sich häufig unscharf, unpräzise und unstrukturiert. Die für die zukünftige wirtschaftliche und gesellschaftliche Entwicklung dringend erforderliche breite Akzeptanz der Informationsgewinnung in weltweiten Datennetzen wird dadurch erschwert. - Die Umstellung der Abfragetechnik auf moderne Windows-Oberflächen und das Internet können diese Einschätzung nur zum Teil relativieren; eine effektive Datenbankrecherche ist für den „Computer-Laien“ nach wie vor ohne angemessene Einarbeitungszeit praktisch unmöglich.

Thematische Abfragen in Datenbanken jeder Art (bibliographische Datenbanken, Faktendatenbanken, etc.) sind nach dem derzeitigen Stand der Technik ausnahmslos an der Verwendung sogenannter Boolescher Operatoren (‘und’, ‘oder’, ‘nicht’) orientiert, mit deren Hilfe komplexere Suchanfragen vom Benutzer strukturiert werden müssen. Die meist in ‘natürlicher Sprache’ präzise formulierbaren Fragen im Zusammenhang thematischer Suchen in Datenbanken, wie z.B.

- Auslandsberichterstattung in den Massenmedien
- emissionsfreie Abfallverbrennungsanlagen in Europa
- der Wald im Unterricht

werden mit Hilfe Boolescher Operatoren auf Strukturen ‘heruntergebrochen’, die mit den ursprünglich syntaktisch-semantisch geordneten Konstruktionen nur noch wenig zu tun haben. Die vom Benutzer der Datenbank mit Ausdrücken wie „ ... in den Massenmedien“, „ ... im Nordosten Frankreichs“, „ ... mit den Mitteln der Spektralanalyse“, „ ... im Unterricht“, usw. explizit formulierte Modifikation eines Themas wird nicht als solche erkannt, sondern muß als eigenständiges Suchwort

bei der Abfrage behandelt werden und würde ausgehend von den oben genannten Beispielen zu folgenden Suchanfragen führen:

- auslandsberichterstattung ***und*** massenmedien
- emissionsfrei$ ***und*** abfallverbrennungsanlage$ ***und*** europa
- wald ***und*** unterricht

Die thematische Frage an die Datenbank wird also vom Retrievalsystem syntaktisch-semantisch darauf reduziert, ob (zufällig) Datensätze vorhanden sind, die die genannten Wörter gleichzeitig enthalten. Dabei kann der Sinn der Ausgangsfrage nicht mehr rekonstruiert werden.

Die Abfrage „wald und unterricht“ bspw. kann aus zwei sehr unterschiedlichen Fragestellungen hervorgegangen sein:

- der Wald im Unterricht oder
- Unterricht im Wald.

Entsprechend gestaltet sich die Präsentation der Suchergebnisse zufällig, unscharf, unpräzise und sehr oft mit der Rückmeldung für den Benutzer „Null Treffer“ oder „zu viele Treffer“.

Weiterführende Retrievaltechniken wie Trunkierung, additional discriminating information (ADI), proximity oder die Verwendung von Relevance-Ranking-Techniken setzen, abgesehen davon, daß sie diese Probleme im allgemeinen auch nicht lösen können, darüberhinaus in der Regel noch detailliertere Kenntnisse des Benutzers voraus.

Unter Einsatz eines speziellen Verfahrens für syntaktisch-semantisches Retrieval werden im Osiris-System multilinguale natürlichsprachliche Datenbankanfragen verarbeitet. Auf der Basis einer deklarativ programmierten Grammatik der natürlichen Sprache, die die Regeln des modellierten Sprachausschnitts berücksichtigt, werden Benutzereingaben syntaktisch und semantisch interpretiert. Dafür werden einschlägige Methoden der semantischen Informationsverarbeitung angewendet (fehlersensitives syntaktisches Parsing, Morphologie, Kompositazerlegung). Die Analyseergebnisse werden unter Berücksichtigung erkannter syntaktisch-semantischer Zusammenhänge auf einer Wissensbasis prozessiert und interpretiert.

Die OSIRIS-Wissensbasis für die Anwendung in der Bibliothek wird vollautomatisch aus den in der OPAC-Datenbank vorhandene Sacherschließungselementen aufgebaut.

Ein speziell entwickelter *OSIRIS-CALCULATOR*, dem ein abstraktes mathematisches Modell klassifikatorischer Systeme zugrunde liegt, gewichtet, bewertet und relationiert Suchergebnisse nach Themen, Einschränkungen und Modifikationen und stellt auf diese Weise auf der Basis der computerlinguistischen Analysen eine syntaktisch-semantische Prozessierung von Suchanfragen auf der OSIRIS-Wissensbasis dar. Die erkannten syntaktisch-semantischen Zusammenhänge gehen dabei nicht durch die ‘sinnleerende’ Reduktion auf Boolesche Verknüpfungen verloren.

Anmerkung: Im Gegensatz zu klassischen Retrievalsystemen führen die Suchanfragen „der Wald im Unterricht“ und „Unterricht im Wald“ in der OSIRIS-Anwendung der Universitätsbibliothek Osnabrück (semantisch-syntaktisch korrekt) zu unterschiedlichen Suchergebnissen.

Die Eingabe zur thematischen Recherche in einem OSIRIS-System ist die Vervollständigung eines auf der Eingabemaske vorgegebenen Satzes. Ohne daß der Benutzer gezwungen wird, sich an eine bestimmte Form der Eingabe zu gewöhnen, wird er dennoch unbewußt zur Formulierung seiner Eingabe mit Hilfe ganz bestimmter syntaktischer Strukturen geleitet.

Für die OSIRIS-Anwendung in einer bibliographischen Datenbank erscheint auf der Maske der Satz „Ich suche Literatur zum Thema ...“, der vom Benutzer für die Recherche vervollständigt werden soll. Dieser Satz ist korrekt nur durch eine Nominalphrase zu ergänzen (z.B. ‘China’), die eventuell eine komplexe innere Struktur haben kann (z.B. ‘China zur Zeit der Kulturrevolution’, ‘Abfallwirtschaft mit dem neuen ...’, ‘Genschers Außenpolitik’, ‘Pädagogik in Frankreich’, etc.). Komplexität und Ambiguität der Eingabe werden so auf eine dem Benutzer natürlich erscheinende Art und Weise auf die Struktur einer Nominalphrase reduziert.

Auf der Basis der Halbsatzergänzung genügt zur Analyse der Benutzereingabe ein relativ kleiner, auf bestimmte syntaktische Phänomene optimierter Parser, der eine effiziente Eingabeverarbeitung garantiert. Ein derartiger Parser besteht aus einer deklarativ programmierten Grammatik, in der die Regeln des modellierten Sprachausschnitts enthalten sind und um den herum sich weitere, die Arbeit des Parsers unterstützende Module gruppieren.

Um komplexe Nominalphrasen, bestehend aus Eigennamen, Einschränkungen und Ergänzungen, etc. verarbeiten zu können, müssen grammatische Informationen zu einzelnen Wörtern der Eingabe verfügbar sein. Solche Informationen sind im *Lexikon* abgelegt, das Angaben über z.B. Wortart, Numerus und Genus enthält. Im Lexikon befindet sich auch eine rudimentäre Semantik der Funktionswörter wie ‘in’, ‘zur Zeit von’, usw., die eine Einschränkung oder Modifikation ausdrücken.

Weiterhin notwendig ist der Einsatz einer *Morphologiekomponente*. Es ist wünschenswert, Eingaben des Benutzers wie ‘Insekten’ und ‘Insekt’ oder auch ‘Massenmedien’ und ‘Massenmedium’ nicht als voneinander unabhängig (d.h. insbesondere mit einem eigenen Lexikoneintrag) zu betrachten. Um die einzelnen morphologisch markierten Formen eines Wortes zueinander in Beziehung zu setzen, ist regelbasiertes Wissen über z.B. Verbflektion notwendig. Mit Hilfe einer Morphologiekomponente können auch Eingaben wie ‘Marktwirtschaft in China’ und ‘chinesische Marktwirtschaft’ zueinander in Bezug gesetzt werden.

Neben der Morphologie-Komponente wird eine Komponente zur *Zerlegung von Komposita* benötigt. Eingaben wie ‘Pädagogikstudium’ können angesichts der im Deutschen ungeheuer produktiven Kompositabildung nicht konsequent als eigenständige Lexikoneinträge behandelt werden, sondern müssen zurückgeführt werden auf ‘Studium der Pädagogik’ oder ‘Studieren von Pädagogik’.

Eine weitere spezielle Komponente stellt die *Behandlung fehlerhafter Eingaben* durch den Benutzer sicher. Schreibfehler bei der Benutzereingabe können aufgrund phonetischer Ähnlichkeit in vielen Fällen erkannt werden. Typische Fehler wie Verdrehen zweier Buchstaben sind durch dieses Modul ebenfalls in eingeschränktem Maße behandelbar.

Die spezielle OSIRIS-Anwendung in der Universitätsbibliothek Osnabrück führt ohne sehr aufwendige, zusätzliche Erschließungsarbeiten zu einer englischsprachigen Oberfläche des Bibliothekskataloges, weil die Bibliothek für die Katalogisierung ihrer Bücher seit ca. 10 Jahren in statistisch hinreichendem Umfang Titeldaten der Library of Congress (mit englischsprachigen Schlagwörtern) verwendet. Die OSIRIS-Wissensbasis konzentiert ebenfalls das englischsprachige Wortmaterial und eröffnet dem Benutzer, bspw. dem ausländischen Wissenschaftler, die Möglichkeit, mit englischsprachigen Suchbegriffen den gesamten Buchbestand der Bibliothek abzufragen.

Die in der OSIRIS-Wissensbasis dargestellten Beziehungen zwischen der lokalen Klassifikation und aus Fremddaten gewonnenen Sacherschließungsmerkmalen eröffnen außerdem effektiven Nutzen im Rahmen eines sog. Computer Aided Indexing (CAI), d.h. dem compupterunterstützten Klassifizieren von Büchern und elektronischen Texten. - Ein weiteres innovatives Merkmal, daß erst in einigen Jahren an Bedeutung gewinnen wird, betrifft die uneingeschränkte Unterstützung der bevorstehenden Rechtschreibreform. Bei klassischen Retrievalsystemen dürften im Zuge der neuen Rechtschreibung in absehbarer Zeit große Probleme auftauchen.

Die Anwendung von OSIRIS in der Universitätsbibliothek Osnabrück zeigt, daß das System zu deutlich qualitativen Verbesserungen sowohl bei der Informationswiedergewinnung als auch bei der Aufbereitung der recherchierten Informationen führt.

Von besonderer wissenschaftspolititscher Bedeutung ist, daß das OSIRIS-System die internationale Kooperation im Wissenschaftsbereich deutlich verbessert. Zum einen werden deutsche Datenbanken durch eine englischsprachige Oberfläche ohne jede Kenntnis des jeweils verwendeten Klassifikationssystems international benutzbar. Zum anderen kann das OSIRIS-System deutschsprachige Suchanfragen automatisch in den Kontext international verbreiteter Klassifikationssysteme transformieren und Recherchen in international bedeutsamen Datenbanken (z.B. Library of Congress für das Bibliothekswesen) deutschsprachig gestalten.