Lexis-Nexis' Freestyle

Natürlichsprachige Suche - More like this!

Insbesondere durch die Suchmaschinen im Internet wurde die Aufmerksamkeit der Information Professionals auf Retrievalmöglichkeiten jenseits der Booleschen Operatoren gelenkt. Auch die kommerziellen Online-Archive entwickelten in den letzten Jahren natürlichsprachige Suchoptionen. Lexis-Nexis erhielt im Laufe des Jahres 1998 zwei Patente für Module automatischer Indexierung erteilt.

Mit natürlichsprachigen Suchen ist gemeint, daß Suchfragen durch die Angabe von Suchwörtern oder durch ganze Sätze ausgedrückt werden, ohne daß irgendwelche Operatoren zu benutzen wären. Das Suchergebnis ist eine nach Wichtigkeit geordnete Liste von Treffern. Der zur Suchfrage bestpassende Datensatz steht oben, der zweitbeste folgt usw. Zwei Ziele werden mit natürlichsprachigen Suchen verfolgt. Zum einen spricht die einfache Suche Endnutzer an, die ohne Kenntnis von mengentheoretischen Operatoren, Abstandsoperatoren, Klammerung und weiteren für sie kryptischen Befehlen nunmehr zu ansprechenden Suchergebnissen kommen. Zum andern ist dieser Retrievalmodus an den professionellen Rechercheur gerichtet. Im Booleschen Retrieval zeigte sich nämlich seit langem, daß ein Recall von 100% und eine Präzision von ebenfalls 100% bei Literaturinformationen unmöglich zu erreichen sind. Eine Steigerung des Recall führt - im theoretischen Modell - zu einer Verringerung der Präzision und umgekehrt. Die erhoffte Summe aus Recall und Präzision erreicht auf keinem Fall die idealen 200, sondern in aller Regel gerademal die Hälfte. Natürlichsprachiges Retrieval soll diesen Mißstand lindern.

In der kommerziellen Informationswirtschaft setzen derzeit bereits einige Hosts auf natürlichsprachiges Retrieval. Zu erinnern ist an DIALOG mit Target, Westlaw mit WIN ("Westlaw Is Natural") und Lexis-Nexis mit Freestyle. Mit Freestyle wollen wir uns hier exemplarisch befassen. Kann ein Endnutzer damit nutzbringend umgehen? Was bringt Freestyle dem Information Professional? Wird die Recall-Präzision-Begrenzung wirklich aufgehoben? Konkret: Bekommen wir mehr relevante Datensätze als beim Booleschen Retrieval? Und: Wird die Nachweismenge präziser, d.h. ärmer an Ballast?

Recall und Präzision

Unser Praxisbericht ist diesmal ausgesprochen theorielastig. Dies liegt in der Natur der Sache. Natürlichsprachiges Retrieval ist nicht selbstverständlich, sondern hat eine Reihe von theoretischen Voraussetzungen. Diese Voraussetzungen muß man kennen, will man in der Praxis erfolgreich mit den entsprechenden Instrumenten umgehen. Als erstes schauen wir uns die Parameter Recall und Präzision genauer an.

Ein "Klassiker" der Informationswissenschaft, Gerald Salton, hat zur Bestimmung der Qualität von Antwortmengen bei Literaturinformation die beiden Aspekte "Recall" und "Präzision" besprochen. (Anmerkung: Das Folgende gilt ausschließlich im Bereich literaturbezogener Informationen, nicht aber bei Fakteninformationen. Dort liegen andere Regelmäßigkeiten vor, die 100% Recall bei gleichzeitig 100% Präzision sehr wohl erlauben.) Wir gehen von drei Mengen aus, wobei

- a =: gefundene relevante Treffer
- b =: nichtrelevante Datensätze, die in der Treffermenge enthalten sind (Ballast)
- c =: relevante Datensätze in der Datenbank, die nicht gefunden wurden (Verlust).

Recall (Vollständigkeit) errechnet sich als Quotient aus der Anzahl der gefundenen relevanten Datensätze und der Gesamtzahl der relevanten Datensätze in der Datenbank:

Recall (in %) = 100 a / a+c.

Präzision ergibt sich als Quotient aus der Anzahl der gefundenen relevanten Datensätze und der Gesamtzahl der gefundenen Datensätze:

Präzision (in %) = 100 a / a+b.

Die Präzision ist - sehen wir von Randunschärfen bei der Bestimmung der Relevanz ab - genau meßbar. Recall ist demgegenüber ein reines Konstrukt, da der Wert c nicht meßbar ist. Woher weiß ich, was ich nicht gefunden habe? Gäbe es Algorithmen, den Verlust zu benennen, würde ich als Rechercheur diese auch einsetzen und keinen Verlust erleiden. Leider gibt es solche Algorithmen außerhalb von experimentellen (überschaubaren) Datenbanken nicht, der Rechercheur wird immer mit Verlust zu kämpfen haben, und es wird der Qualitätsmessende nie wissen, wie groß c tatsächlich ist.

Das Konstrukt "Recall" ist in Verbindung mit der "Präzision" jedoch in einem Denkmodell nützlich. Versuchen wir im Booleschen Retrieval, die Recallquote zu erhöhen (etwa durch mit ODER angehängte weitere Suchargumente), so sinkt gleichzeitig die Quote der Präzision, d.h. wir erkaufen uns einen besseren Grad an Vollständigkeit mit einem schlechteren Grad an Genauigkeit, sprich mit Ballast. Weisen wir alle Datensätze einer Datenbank nach, so erwischen wir auch alle relevanten Datensätze. Der Recall-Wert liegt bei idealen 100%. Der Ballast ist riesig; die Präzision liegt bei etwa Null, das Rechercheergebnis ist unbrauchbar.

Versuchen wir umgekehrt, die Genauigkeit zu erhöhen (nunmehr z.B. durch mit UND angehängte weitere Suchargumente), also Ballast möglichst völlig zu vermeiden, verringern wir die Vollständigkeit. Ein völlig ballastfreies Retrievalergebnis könnte ebenso unbrauchbar sein, da vielzuviele Informationen gar nicht aufscheinen. In diesem Wechselspiel von Vollständigkeit und Ballast muß der Recherchierende ein optimales Ergebnis herstellen - ein Verfahren, das sich algorithmischen Verfahren entzieht und dessen Güte entscheidend von den Erfahrungen und der "Kunst" des Rechercheurs abhängt.

Ein optimales Retrievalergebnis läge sowohl im Recall und in der Präzision bei jeweils 100%. Angesichts des eben geschilderten Zusammenhangs zwischen Recall und Präzision wird dieses Ideal in der Abbildung 1 als der kaum erreichbare "heilige Gral" der Rechercheure eingezeichnet.

Die beiden Booleschen Operatoren UND und ODER werden zwecks Orientierung am "heiligen Gral" verfeinert. Der erste Weg geht vom UND aus und verstärkt es zu Abstandsoperatoren. Der zweite Weg führt uns zu den natürlichsprachigen Suchen, die zu großen Teilen eine Variation des Booleschen ODER darstellen.

Vertreter natürlichsprachiger Suchen wie Ross Evans behaupten nun, daß - zusätzlich zu im Booleschen Retrieval entdeckten Literaturstellen - weitere relevante Treffer gefunden werden, daß wir uns also mit natürlichsprachigem Retrieval dem "heiligen Gral" in der Tat ein Stück nähern (siehe *Abbildung 1*).

Aspekte automatischer Inhaltserschließung

Wir wollen uns einen knappen Überblick darüber verschaffen, mit welchen Mitteln natürlichsprachige Systeme arbeiten. Allen Systemen gemein ist die Annahme, daß das System zur automatischen Erschließung elektronisch vorliegender Informationsinhalte und zur Erschließung der Suchanfragen geeignet ist. Ziel ist es, solche Informationsinhalte zu finden, die der Suchanfrage möglichst ähnlich sind.

Die theoretischen Grundlagen automatischer Inhaltserschließung lassen sich in drei Aspekte einteilen, die allein oder in Kombinationen in konkreten Systemen eingesetzt werden:

- linguistische Aspekte (Stoppwörter, Wortstammanalysen, Phrasenerkennung, Pronomina-Analysen)
- statistische Aspekte (Worthäufigkeiten, Rangordnungsalgorithmen)
- ordnungstheoretische Aspekte (Deskriptor-, Notations- bzw. Schlagwortvergabe).

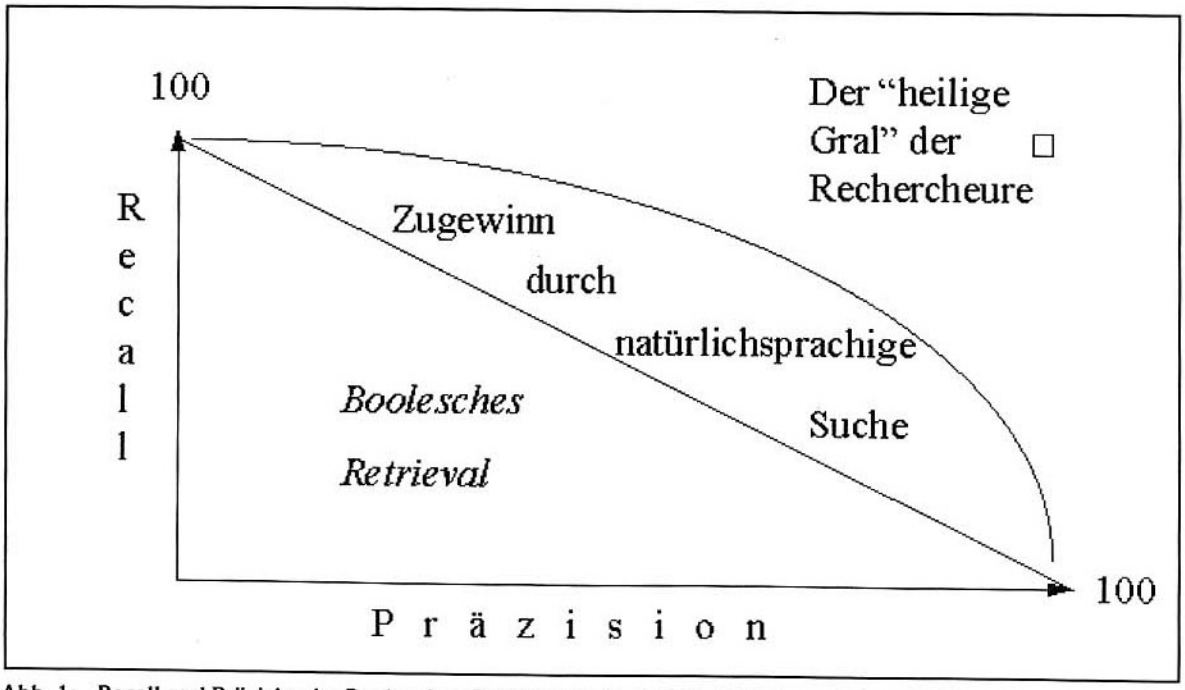

Abb. 1: Recall und Präzision im Booleschen Retrieval und bei der natürlichsprachigen Suche nach Literaturinformationen *(Quelle:* Evans 1994, 124)

Informationslinguistik

Informationslinguistische Methoden, die derzeit in Retrievalsystemen Einsatz finden, haben vorwiegend die Aufgabe, die Bildung von Sucheinstiegen in einen Text vorzubereiten. Es geht darum, nicht sinntragende Wörter zu eliminieren, grammatische Flexionsformen auf eine Grundform zu bringen, aus mehreren Termen bestehende Phrasen zu erkennen sowie Pronomina den jeweiligen Nomen zuzuordnen.

Triviale Voraussetzung aller weiterer Schritte ist die Isolation einzelner Wörter. Dies sind Zeichenfolgen, die zwischen zwei Leerzeichen, Satzzeichen o.ä. stehen.

Elimination von Stoppwörtern. Während mengentheoretisch vorgehende Retrievalsysteme mit rund zehn Stoppwörtern auskommen ("the", "a", "an", "but" usw.), erhöhen natürlichsprachige Systeme deren Anzahl auf bis zu 500. Eingeschlossen sind Wörter, von denen erwartet wird, daß sie normalerweise keine eigenständigen thematischen Passagen ausdrücken. Beispiele sind "meanwhile", "nevertheless" oder "myself". Problematisch ist die oftmals durchgeführte Elimination von Pronomina, kommt es doch so zu Verzerrungen bei informationsstatistischen Auszählungen. Pronomina sollten zunächst markiert und für informationsstatistische Zählungen verwendet werden, erst danach sind sie für die Suche zu eliminieren. (Wir werden bei den Pronomina-Analysen hierauf zurückkommen.)

Wortstammanalysen. Grammatikalische Flexionsformen werden auf deren Wortstamm reduziert. In der englischen Sprache reicht es in der Regel aus, bei Beachtung einiger Regeln die Endungen (über eine vorgegebene Suffixliste) zu tilgen. Ist z.B. "-ing" in der Liste enthalten, wird das Wort "ringing" auf den Stamm "ring" reduziert (aber durch eine an der Länge orientierten Regel nicht auf "r"). Alle Varianten eines Wortstammes gelten dabei als dasselbe Wort.

Phrasenerkennung. Eine Phrase ist ein Ausdruck, der aus mehreren einzelnen Wörtern besteht. Hier gilt nicht das einzelne Wort (oder dessen Wortstamm) als Schlagwort, sondern die Phrase als Ganzes. Das System muß beispielsweise in der Lage sein, "Information Retrieval", den Körperschaftsnamen "Fachhochschule Köln" oder den Eigennamen "Wolfgang Clement" als eine Einheit zu erkennen. Dies geschieht entweder durch den Abgleich mit vorhandenen Listen oder durch eine Analyse gemeinsamen Auftretens in den Dokumenten.

Pronomina-Analysen. Die sinntragenden Wortstämme sowie die Phrasen werden, wie wir gleich sehen werden, als Zählbasis für statistische Berechnungen verwendet. Es ist demnach die Markierung jedes Vorkommens in einem Text nötig. Pronomina, die an die Stelle ihrer Nomen rücken, müssen dabei entsprechend beachtet werden. Betrachten wir den Satz: "The president has a girl friend, but he doesn't love her", so muß das "he" dem Wort "president" und "her" der Phrase "girl friend" zugeordnet werden.

Informationsstatistik

Informationsstatistik zählt Wörter sowie Datensätze und setzt die ausgezählten Werte in bestimmte Relationen. Was jeweils als ein Wort gilt, hängt vom Umfang der eingesetzten informationslinguistischen Methoden ab. So gibt es zum Teil riesige Unterschiede, ob alle Flexionsformen einzeln gezählt werden, ob nur die Wortstämme oder ob Wortstämme und zugehörige Pronomina gezählt werden. Je elaborierter die informationslinguistischen Verfahren arbeiten, desto bessere Resultate wird auch die Informationsstatistik erbringen. Hauptziel der Informationsstatistik sind Algorithmen der Rangordnung von Datensätzen nach Relevanz ("relevance ranking").

Einfache Zählungen von Worthäufigkeiten. Alle Wörter in einem Text und - als Summe über die Texte - alle Wörter in einer Datenbasis werden gezählt. Hierbei wird jedem Wort eines Textes dessen Häufigkeit im Text zugeordnet. Zudem wird ausgezählt, in wievielen Texten das entsprechende Wort überhaupt vorkommt. Die so ermittelten Summen sind die Rohwerte für alle weiteren Berechnungen.

Einige Systeme arbeiten auf der Basis der Anzahl der Dokumente, in denen ein Wort vorkommt, mit einem Schwellenwert. Ab einem gewissen Wert werden die so markierten Hochfrequenzbegriffe eliminiert, da sie für Retrievalzwecke nicht genügend trennscharf sind.

Dokumentspezifische Wortgewichtung. Eine einfache Zählung der Worthäufigkeit in einem Dokument eignet sich nicht für informationsstatistische Zwecke, da längere Texte methodisch bevorzugt würden. Vielmehr arbeitet man mit der relativen Häufigkeit von Textwörtern, d.h. mit dem Quotienten aus der Häufigkeit des Wortes und der Gesamtmenge der Wörter im Text.

Gewichtung nach Position im Text. Die dokumentspezifische Wortgewichtung wird verfeinert, indem die Stellung in Textteilen Berücksichtigung findet. So kann ein Text in mehrere Teile zerlegt werden. Wenn man annimmt, daß Terme am Textanfang wichtiger sind als solche in der Mitte, so wird dies durch geeignete Gewichtungsfaktoren berücksichtigt. Ein Vorkommen im ersten Teil kann etwa mit 1,5 bewertet werden, ein Vorkommen in einem mittleren Abschnitt mit 0,8 usw. Eine Variante dieses Verfahrens arbeitet mit den dokumentarischen Beschreibungen der Texte. Die Termgewichtung orientiert sich hier am Auftreten in bestimmten Feldern. Ein Vorkommen im Sachtitel wird z.B. höher gewichtet als ein Vorkommen im Abstract.

Inverse Dokumenthäufigkeit. Während ein Textwort im Kontext eines Textes um so wichtiger ist, je häufiger es vorkommt, gilt in Beziehung zur Gesamtdatenbank die umgekehrte Proportion. Ein Wort ist demnach um so wichtiger, je weniger Dokumente dazu in der Datenbank vorhanden sind. Um die Spannweite der Gewichtungswerte nicht allzu groß werden zu lassen, arbeitet man bei der Berechnung der inversen Dokumenthäufigkeit mit logarithmischen Werten. Die klassische Berechnungsformel lautet:

$$IDF(i) = (\log_2 N / n) + 1,$$

wobei IDF(i) die inverse Dokumenthäufigkeit des Wortes i, N die Gesamtanzahl der Datensätze in der Datenbank und n die Anzahl der Datensätze, in denen i vorkommt, ist. Der Wert des IDF ist - für jedes Wort - in ständiger Bewegung, insofern die Datenbank wächst und so auf jedem Fall N ändert.

Rangordnungen. Mit der IDF ist ein weiterer Gewichtungsfaktor gefunden. Für jedes Wort eines Textes ergibt sich durch Multiplikation der dokumentspezifischen Wortgewichtung (mit oder ohne Berücksichtigung der Positionen im Text)

mit seiner inversen Dokumenthäufigkeit ein Gewichtungswert. Betrachtet man mehrere Wörter, so addieren sich die einzelnen Gewichtungswerte pro Dokument. Diese Summe bildet die Basis für Sortierungen nach Relevanz.

Ordnungstheorie

Im Rahmen der Ordnungstheorie geht es um die Zuordnung von kontrolliertem Vokabular (Deskriptoren, Notationen) bzw. von Schlagworten zu einem Datensatz. Deskriptoren kann man nur dann verwenden, wenn ein Thesaurus vorliegt, Notationen nur im Rahmen eines Klassifikationssystems. Die Schlagwortvergabe arbeitet ausschließlich mit dem in den Texten vorhandenen Termmaterial. Diese Methoden schließen einander nicht aus, können also ggf. gemeinsam eingesetzt werden.

Automatische Deskriptor- oder Notationsvergabe beim Einsatz einer Dokumentationssprache. Eine Dokumentationssprache (Thesaurus oder Klassifikationssystem), die zur automatischen Indexierung herangezogen wird, muß über besonders ausgefeilte Synonymie-Relationen verfügen. Die Gewichtungswerte aller einem Text zugeordneten Schlagworte werden mit einem Schwellenwert verglichen. Für die verbleibenden wichtigsten Terme geschieht ein Abgleich mit der Liste der Deskriptoren und ihrer Synonyme. Ein Deskriptor wird dann zugeteilt, wenn er (bzw. sein Wortstamm) selber oder wenn einer seiner Synonyme vorkommt. Analog kann bei Klassifikationen vorgegangen werden. Diese Variante der automatischen Indexierung fundiert ein "normales" Boolesches Retrieval mit kontrolliertem Vokabular.

Automatische Schlagwortvergabe. Alle nach den informationslinguistischen und -statistischen Methoden markierten Textwörter sind Kandidaten für Schlagworte. Gemäß Gerald Salton kann man jedes Schlagwort als Dimension verstehen. Eine Datenbank mit n verschiedenen Schlagworten spannt demnach einen n-dimensionalen Raum auf. Die Repräsentation eines Dokuments ist nach Salton derjenige Vektor, der durch alle jeweils vergebenen Schlagworte verläuft. (Nach diesem Grundgedanken wurde Saltons berühmtes System SMART, Salton's Magical Automated Retrieval Technique, kreiert.)

Da die Terme durch die statistischen Verfahren mit einem dokumentspezifischen numerischen Wert versehen sind, bekommt unser Vektor auf jeder Dimension einen konkreten Wert zugeordnet. Es ist hierbei möglich, mit einem Schwellenwert zu arbeiten. Ein Vektor läuft nur dann durch eine Dimension, wenn der Schwellenwert überschritten wird.

Durch die Analyse der Vektoren läßt sich ein "statistischer Thesaurus" aufbauen. Dimensionen, die von mehreren Vektoren gemeinsam berührt werden, gelten dabei als "verwandt" oder "ähnlich". Hierbei wird in der Regel mit mehreren Schwellenwerten gearbeitet. Gefordert ist ein hoher Gewichtungswert der einzelnen Textwörter sowie eine möglichst große Zahl gemeinsamen Auftretens unterschiedlicher Schlagworte innerhalb der Vektoren.

Nach dem Gesagten ist die Bestimmung der automatischen Indexierung bei der Variante über Schlagworte recht einfach. Sie ist ein Abgleich des Vektors, der die Suchanfrage repräsentiert, mit den Vektoren der gesamten Datenbank. Ausgegeben werden die Dokumente derjenigen Vektoren, die dem Suchvektor am ähnlichsten sind.

Freestyle in der Praxis

Genug der Theorie! Wir wenden uns der Praxis zu. Lexis-Nexis arbeitet mit zwei Retrievaloptionen: mengentheoretisches Retrieval (eingeleitet mit .bool) und natürlichsprachiges Retrieval (eingeleitet mit .fr). Die Suche bei Freestyle wird durch einen natürlichsprachigen englischen Satz formuliert. Freestyle analysiert hierbei ausschließlich die vorkommenden Wörter, nicht deren grammatikalischen Zusammenhang. (Man kann also auch mit einem nicht so ausgefeilten Englisch arbeiten, einfache Wortlisten reichen auch aus.) Phrasen sollten im Fragesatz mit Anführungszeichen gekennzeichnet werden, da Freestyle nicht alle Phrasen erkennt. Verwandte Begriffe, Synonyme oder Quasisynonyme werden einem Suchwort in Klammern hinzugefügt. Da Fragmentierungen nicht möglich sind, müssen hier auch Formulierungsvarianten angegeben werden.

Freestyle bietet fünf Optionen zur Bearbeitung einer Suchfrage an (siehe Abbildung 2). Option (1) erlaubt das Streichen und Hinzufügen von Termen. Durch (2) können solche Wörter ausgezeichnet werden, die im Text notwendig vorkommen müssen. Hier kommt durch eine Hintertür das mengentheoretische UND wieder ins Spiel. Option (3) läßt gewisse Einschränkungen zu, etwa zu bestimmten Datumsangaben oder auf Terme, die im Titel vorkommen müssen. (Auch hier liegt wiederum eine Schnittmengenbildung vor.) Wird nichts eingesetzt, so läuft die Suche stets über den gesamten File. (4) führt zu Synonymenwörterbüchern sowie zu "verwandten Begriffen". Aus dem zum Teil reichhaltigen Termangebot wird mit weiteren Wörtern die Suche verfeinert. Option (5) legt die Anzahl der auszugebenden Datensätze fest. Standardeinstellung ist 25; die Menge ist bis auf 1.000 erweiterbar.

Unsere Beispielsuche in *Abbildung 2* wurde in der "Germany Library" im File "ALLNWS" (All News) durchgeführt.

```
SEARCH DESCRIPTION:
WHAT OPINIONS FOR THE FUTURE OF GARZWEILER AND HAMBACH
HAVE RHEINBRAUN AND WOLFGANG CLEMENT?

Press ENTER to start search.
<=1>  Edit Search Description
<=2>  Enter/edit Mandatory Terms
<=3>  Enter/edit Restrictions (e.g., date)
<=4>  Synonyms and Related Concepts
<=5>  Change number of documents  Current setting: 50

For further explanation, press the H key (for HELP) and
then the ENTER key.

FREESTYLE(TM) SEARCH OPTIONS
```

Abb. 2: Freestyle: Optionen zur Bearbeitung einer Suchfrage

```
Synonyms for: OPINIONS   (Page 1 of 2) Press ENTER for next page
Enter synonym numbers to include in search and press ENTER
(e.g. 1,2,3-4)

<=1> Return to Search Options  <=2> Return to Term Selection

---------------------------------- Belief ------------------

 1 assumption              2 attitude            3 conclusio
 4 conjecture              5 consideration       6 conviction
 7 determination           8 estimate            9 estimation
10 evaluation             11 feeling            12 guess
13 hypothesis             14 idea               15 impression
16 judgment               17 notion             18 outlook
19 perspective            20 persuasion         21 point of view
22 position               23 posture            24 preconception
25 presumption            26 presupposition     27 reaction
28 reflection             29 sentiment          30 stance
31 stand                  32 suppositio         33 surmise
34 suspicion              35 theory             36 thesis
37 thinking               38 thought            39 view

For further explanation, press the H key (for HELP) and
then the ENTER key.
                                   SYNONYM SELECTION
```

Abb. 3: Freestyle: Wörterbuchunterstützung

Zur Veranschaulichung der Mächtigkeit der Wörterbücher sei hier ein Ausschnitt der angebotenen Varianten zu "Opinions" abgedruckt (*Abbildung 3*). Quelle der Wörterbücher sind "The Legal Thesaurus" von William C. Burton sowie "Merriam Webster's Collegiate Thesaurus". Durch Angabe der laufenden Nummern werden die Terme der Suchfrage hinzugefügt.

Die Sortierung der gefundenen Datensätze erfolgt nach Relevanz. Es besteht die Möglichkeit, die Trefferliste - wie gewohnt - nach der Zeit neu zu sortieren. Die Dokumentanzeige geschieht sowohl durch die üblichen Optionen (bibliographische Angaben, Keyword in Context oder Volltext) als auch zusätzlich durch eine Option "SuperKWIC", die den entsprechend zur Suchfrage wichtigsten Textteil anzeigt.

Auch nach der Anzeige der Treffer sind Modifikationen der Suchanfrage möglich. Dies geschieht durch "MODIFY" bei der Dokumentanzeige und bewirkt ein erneutes Aufrufen des Optionenmenüs (wie in *Abbildung 2*).

Boole und Freestyle in Kombination

Das Retrievalsystem von Lexis-Nexis bietet an, die durch Freestyle gefundene Treffermenge mittels mengentheoretischer Operatoren weiter zu bearbeiten. Diese Suchverfeinerung wird durch "FOCUS" eingeleitet. Hier lassen sich ausgesprochen interessante Suchstrategien abarbeiten. In einem ersten Schritt grenzt man durch Freestyle eine umfangreiche Menge von Datensätzen auf die relevantesten Dokumente ein. Im zweiten Schritt arbeitet man in dieser Teilmenge mit den gewohnten Booleschen Operatoren. Geschickt eingesetzt, läßt sich die Präzision einer Treffermenge so enorm erhöhen.

Der umgekehrte Weg ist ebenfalls möglich. Zunächst wird mengentheoretisch gesucht, danach mittels relevance ranking sortiert. Möglich ist diese Retrievalstrategie allerdings nur bei Treffermengen bis maximal 1.000 Datensätzen.

Nach Abschluß der Booleschen Suche leitet man durch den Befehl .rank die Sortierung nach Wichtigkeit ein. Voraussetzung für das relevance ranking ist eine thematische Suche. Eine Suche nach Formalia (etwa "date is 10/9/1998") läßt keine Sortierung nach Relevanz zu.

Gibt es bei großen Treffermengen überhaupt Alternativen zum relevance ranking? Eine Beipielsuche bei ALLNWS (Germany Library) nach "Garzweiler" brachte 1.080 Treffer. Freestyle listet (bei entsprechender Einstellung) die ersten 1.000 Datensätze nach Wichtigkeit auf und erlaubt (auch mengentheoretische) Verfeinerungen. Eine Beschränkung auf Titel (mittels "headline(Garzweiler)"; 215 Treffer) bringt uns kaum weiter, da Titel oftmals recht ausdruckslos sind. Recht interessant verlaufen Recherchen mit mehrfach eingesetzten Abstandsoperatoren. Wir probierten es mit "Garzweiler w/25 Garzweiler w/25 Garzweiler" und erhielten 33 hochrelevante Texte. Das Finden des "richtigen" Abstandes ist hierbei allerdings ausgesprochen vage, so daß auch dies keine Alternative (wohl aber eine Ergänzung) zu Freestyle ist.

Automatische Indexierung bei Lexis-Nexis

Wie arbeitet Freestyle? Automatisch indexiert werden sowohl die jeweilige Suchfrage als auch der Gesamtbestand an Datensätzen.

Stoppwörter werden eliminiert (in unserem Beispiel aus Abbildung 2 die Terme "what", "the", "of", "and" und "have"). Terme, die in der Datenbank allzu häufig vorkommen, die also angesichts ihres Wertes der inversen Dokumenthäufigkeit einen sehr kleinen Gewichtungswert haben, werden ebenfalls ausgeschlossen (bei uns waren dies "for" und "future"). Unter .why wird diese Eliminationsart dem Benutzer mitgeteilt.

Über die verbleibenden Wörtern wird eine Suche nach Phrasen eingeleitet. Gesucht wird nach Textklumpen, also nach solchen Wortfolgen, die zwischen Stoppwörtern oder Satzzeichen stehen. Zusätzlich werden Eigennamen von Personen durch eine Liste (englischer) Vornamen, Namen von Firmen durch spezifische Bezeichnungen (etwa "Ltd."), Namen von Organisationen durch Zusätze wie "University" oder "Foundation"

erkannt. Lexis-Nexis hat die Algorithmen und Listen durch ein Patent (Lu u.a. 1998) schützen lassen. Phrasen werden dem Nutzer durch Anführungsstriche gekennzeichnet.

Letztlich wird eine Art rudimentäre Wortstammanalyse durchgeführt. Bearbeitet werden jedoch ausschließlich regelmäßige Plural- und Possessivformen. "City" findet demnach "City", "Cities", "City's" und "Cities'". Unregelmäßige Formen (wie z.B. "Child") werden nicht erkannt. Einige Äquivalente, vorrangig Abkürzungen, werden einander zugeordnet, so z.B. "Ala" und "Alabama", "19th" und "Nineteenth" oder "Oct" und "October". Freestyle verfügt über keinen Algorithmus zur Analyse der Pronomina.

Alle in der Suchfrage erkannten Terme werden in einer Schlagwortliste gesammelt und bilden so den Suchvektor.

Informationsstatistisch arbeitet Freestyle sowohl mit der inversen Dokumenthäufigkeit als auch mit der dokumentspezifischen Wortgewichtung. Der Gewichtungswert wird auf ein Intervall abgebildet, dessen Werte zwischen 1 und 100 liegen. Die normierten Gewichtungen für die einzelnen Schlagworte, die durch den Suchvektor vorgegeben sind, werden datensatzweise addiert. Hierdurch wird die Rangordnung hergestellt.

Ein Konkurrenzvorteil von Westlaw gegenüber Lexis-Nexis ist die (intellektuelle) Inhaltserschließung der Rechtstexte bei Westlaw. Lexis-Nexis arbeitet daran (und hat die Bemühungen durch ein weiteres Patent, Mehrle 1998, schützen lassen), Rechtsdokumente automatisch zu klassifizieren.

.where und .why: Dem System über die Schulter schauen

	Documents Retrieved	Documents Matched	Term Importance (0-100)
HAMBACH	8	373	29
RHEINBRAUN	49	675	25
GARZWEILER	50	1080	23
CLEMENT	48	4013	15
OPINIONS	1	13003	9
WOLFGANG	46	89679	1
FUTURE	--	--	--
FOR	--	--	--

Abb. 5: .WHY: Welche Gewichtungswerte erhalten die Suchargumente?

Den Abgleich des Suchvektors mit den Vektoren der Dokumente kann der Rechercheur partiell verfolgen. Hierzu arbeitet man mit den Befehlen .where und .why.

Der WHERE-Bildschirm (*Abbildung 4*) zeigt die Schlagworte, mit denen die Suchfrage durch Freestyle beschrieben wurde, und deren Vorkommen in den ersten 25 Dokumenten. Hat der Nutzer die Anzahl der Dokumente auf einen höheren Wert als 25 gestellt, so werden auch die folgenden Datensätze - immer in 25er Blöcken - auf weiteren Bildschirmen angezeigt. Um ein spezielles Dokument anzusehen, muß man seine Nummer eingeben.

Der WHY-Bildschirm (*Abbildung 5*) listet ebenfalls alle Schlagworte der Freestyle-Suche auf. Genannt werden auch diejenigen Terme, die wegen eines zu niedrigen Gewichtungswertes keine Berücksichtigung beim Retrieval fanden. Der Rechercheur bekommt zu jedem "guten" Schlagwort drei Angaben:

- die Anzahl der Dokumente zu dem Schlagwort in der Treffermenge
- die Anzahl der Dokumente zu dem Schlagwort in der Datenbank
- der auf das Intervall [0,100] normierte Gewichtungswert der inversen Dokumenthäufigkeit (IDF).

Die Schlagworte sind nach Wichtigkeit geordnet. In unserer Suche kommt nur das Schlagwort "Garzweiler" in allen 50 Datensätzen vor. Jedes Vorkommen dieses Wortes in einem Text wird mit dem Faktor 23 gewichtet. "Hambach", in der Gesamtdatenbank viel weniger besprochen als "Garzweiler", erhält entsprechend einen größeren Gewichtungswert. Das Wort "Hambach" wird mit 29 gewichtet. Ein Wort wie "Wolfgang", knapp 90.000 mal in der Datenbank, fällt kaum ins Gewicht. Ein Auftreten von "Hambach" ist im Schnitt genauso wichtig wie 29mal "Wolfgang".

Die Informationen aus den WHERE- und WHY-Bildschirmen zeigen, wie Freestyle mit einer Suchfrage umgegangen ist. Damit klärt sich für den Suchenden, "was überhaupt abläuft". Man kann aber hier auch in das Geschehen eingreifen, indem man etwa gewisse Suchterme als "mandatory" kennzeichnet, einige Terme in der Suchfrage tilgt, Synonyme löscht oder hinzufügt usw. Die natürlichsprachige Suche läßt sich nämlich durchaus optimieren, indem man mit Formulierungsvarianten arbeitet und durch .where und .why seine Frage verbessert.

```
                                1                   2
                1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
HAMBACH         * *   * *           *       *       *         *
RHEINBRAUN      * * * * * * * * * * * * * * * * * * * * * * * * *
GARZWEILER      * * * * * * * * * * * * * * * * * * * * * * * * *
CLEMENT         * * * * * * * * * * * * * * * * * * * * * * * *
OPINIONS
WOLFGANG        * * * * * * * * * * * * * * *   * * * * * * * *
```

Abb. 4: .WHERE: Wo kommen die Suchargumente vor?

Assoziatives Retrieval: *More*

Gesetzt den Fall, der Suchende hat einen Datensatz gefunden, der voll zutreffend ist, dann kann er mit diesem Dokument weitersuchen. Der Befehl "More like this" besagt: "Suche mir Datensätze, die meinem Muster so ähnlich wie möglich sind"! Lexis-Nexis öffnet sich damit einer weiteren Retrievalform, dem assoziativen Retrieval. Das Suchargument ist keine mengentheoretische Verknüpfung, auch kein natürlichsprachiger Satz, sondern ein ganzer Text, der als Muster fungiert.

In *Abbildung 6* gehen wir von einem Patent aus. Wie man an sein Muster kommt, ist irrelevant. Es kann durch Boolesches oder natürlichsprachiges Retrieval gefunden worden sein. Es kann auch schon längst bekannt sein; man ruft es in diesem Fall bei Lexis-Nexis auf. Das assoziative Retrieval wird durch den Befehl .more initiiert. Das Musterdokument muß genügend lang sein, sonst findet das System keine Suchbegriffe.

```
5,144,317
Sep. 1, 1992
Method of determining mining progress
in open cast mining by means of satel-
lite geodesy
INVENTOR: Duddek, Herbert, Bergheim,
Federal Republic of Germany
Klemmer, Wilfried, Dahlem, Federal
Republic of Germany
Koeppen, Herbert, Bergheim, Federal
Republic of Germany
ASSIGNEE-AT-ISSUE: Rheinbraun Akti-
engesellschaft, Koln, Federal Republic
of Germany (03)
... (gekürzt) ...

.more

Search Description:
RECEIVER, BUCKET, WHEEL, EXTRACTION,
SATELLITE, DEPOSIT, MINING, EXCAVATOR,
APPARATUS, CO-ORDINATE, SIGNAL, JIB,
"Pat. No. 4675684"

LEXIS searched for the following cita-
tion and its parallels:
"Pat. No. 4675684"
```

Abb. 6: MORE: Automatische Kreation von Suchargumenten für Freestyle

```
rel(gerhard schroeder) and (rel(helmut kohl))

TOPIC: "GERHARD SCHROEDER", "HELMUT KOHL"
Enter concept numbers of interest and press ENTER
(e.g. 1,2,3-4)

 <=1>  Exit Related concepts
-------------------------------------------------------
  1 Germany                          2 chancellor
  3 coalition                        4 reform
  5 Christian Democratic Union       6 poll
  7 SOCIAL DEMOCRATIC PARTY          8 campaign
  9 employer                         10 grand coalition
 11 politician                       12 victory
 13 vote                             14 voter
 15 welfare                          16 German
 17 Greens                           18 Social Democrats
```

Abb. 7: Related Concepts

MORE arbeitet in zwei Varianten, erstens mit Schlagworten, die automatisch dem Muster entnommen werden, und zweitens mit Zitationen.

Die zitatenanalytische Variante von MORE ist in solchen Fällen möglich, wo in standardisierter Form zitiert wird, also z.B. bei Patentnummern oder Gerichtsurteilen. Gesucht wird nach solchen Patenten, Fällen usw., deren Zitationen am meisten mit den Zitationen des Modelldokuments übereinstimmen. Das Maß an denselben Zitationen gilt hierbei ein Ähnlichkeitsmaß der betreffenden Texte.

Die Schlagwortvariante von MORE arbeitet mit Freestyle, d.h. es werden aus dem Musterdokument automatisch Schlagworte selektiert. Wie bei Freestyle, so können auch hier die Suchwörter bearbeitet werden.

Das assoziative Retrieval verläuft in vielen Fällen zufriedenstellend, insofern interessante neue Dokumente gefunden werden. Grundvoraussetzung ist, daß mit einem ideal passenden Mustertext gearbeitet wird.

Verwandte Begriffe: Anwendung des statistischen Thesaurus

Häufig in einem Vektor gemeinsam vorkommende Wörter werden bei Lexis-Nexis durch einen eigenen Befehl "Related Concepts" aufgelistet. Mittels "rel(Suchterm)" im Booleschen Retrieval oder durch Option (4) im Freestyle-Bearbeitungsbildschirm (Abbildung 2) erhält der Rechercheur zu einzelnen Suchwörtern oder Suchphrasen eine Liste von (mehr oder weniger) verwandten Begriffen. Da hier ausschließlich über Zählmaße gearbeitet wird, darf nicht mit der Qualität einer Assoziationsrelation eines Thesaurus gerechnet werden, sondern durchaus mit zum Teil passenden, zum Teil aber völlig inadäquaten Wörtern. Dafür arbeitet die "Related Concepts"-Software auch mit mehreren Suchwörtern. Sucht man, wie wir in *Abbildung 7*, mit zwei Phrasen, so erhält man diejenigen Terme, die mit beiden Suchargumenten häufig zusammen auftreten (und die zusätzlich in den Texten häufig vorkommen). Die empirischen ordnungstheoretischen Analysen bringen interessante und praktisch brauchbare Resultate.

Die Liste der gemeinsamen verwandten Begriffe zu "Gerhard Schröder" und "Helmut Kohl" (recherchiert zur Zeit der deutschen Bundestagswahl 1998) enthält erwartungsgemäß "chancellor" (Term 2), "Germany"/ "German" (Terme 1 und 16) und die Namen ihrer beiden Parteien (Terme 5 und 7), aber auch Überraschungen wie "grand coalition" (Term 10) oder die "Greens" (Term 17).

Die Wörtersammlung wird als Vorschlagsliste verstanden. Der Nutzer selektiert einige Begriffe und übernimmt diese in seine (mengentheoretische oder natürlichsprachige) Suche. Sucht man im Rahmen der Booleschen Logik, so ist die Wahl des Operators freigestellt. Dies ist ein weiterer Unterschied zur Assoziationsrelation eines Thesaurus, die in Richtung eines Booleschen ODER führt.

Lexis-Nexis führt zwei "statistische Thesaur", einen für den Rechtsbereich und einen zweiten für den News-Teil.

Bewertung

Kommen wir mit Freestyle dem "heiligen Gral" der Rechercheure näher? Allerdings: ja, wir finden Datensätze, die im Booleschen Retrieval nicht gezeigt werden, erhöhen demnach den Recall, und haben gleich mehrere Möglichkeiten, durch relevance ranking und den kombinierten Einsatz von Freestyle und Boolescher Suche die Präzision der Ergebnismenge zu erhöhen. Das assoziative Retrieval mit Musterdatensätzen (More) eröffnet neue Retrievalperspektiven. Zudem kann die Idee des statistischen Thesaurus (mit den Related Concepts) nur begrüßt werden.

Freestyle und More ermöglicht dem Laien, ohne fundiertes informationspraktisches Wissen Ergebnisse zu bekommen. Aber erst in der Anwendung eines Information Professional unter Ausnutzung aller Optionen zeigt sich die Mächtigkeit dieses Retrievalsystems richtig.

Eine Freestyle-Suche kostet genausoviel wie eine Boolesche Suche, dessen Ergebnisse es ergänzt. Damit verteuert sich die Lexis-Nexis-Recherche, aber wir meinen: Es lohnt sich. Fünf Punkte von fünf möglichen für das Preis-Leistungsverhältnis. (Hiermit wird allerdings nichts über die Preise von Lexis-Nexis generell ausgesagt.)

Freestyle ist leicht zu erlernen. Zum Teil erscheint es etwas umständlich, wenn z.B. die "Mandatory Terms" in einem eigenen Arbeitsgang eingegeben werden. Störend ist das Fehlen von Fragmentierungsmöglichkeiten. Trotzdem: vier Punkte für den Bedienungskomfort.

Die informationslinguistischen, -statistischen und ordnungstheoretischen Vorgaben bei kommerziellen Systemen natürlichsprachiger Suchen - und dies gilt nicht nur für Freestyle, sondern auch für Target, WIN und weitere Systeme - sind bei weitem noch nicht ausgereizt. Wortstammanalysen können besser werden, die Analyse der Pronomina muß in Angriff genommen werden. Bei fachlich einheitlichem Material wäre ein Thesaurus oder Klassifikationssystem, dessen Deskriptoren und Notationen auf der Basis automatischer Indexierung den Dokumenten zugeordnet würden, durchaus hilfreich. (Profound kann dies, und das ist ein großer Wettbewerbsvorsprung. Lexis-Nexis arbeitet für den Lexis-Teil daran.) Wir wollen die theoretischen Vorgaben durch unsere Bewertungsdimension "Qualität der Inhalte" ausdrücken. (Um andere Inhalte geht es ja nicht, sondern nur um ein anderes Verfahren, an die Inhalte heranzukommen.) In dieser Dimension kann noch viel geleistet werden, deshalb nur drei Punkte.

Trotz der durchaus vorhandenen Ausbaumöglichkeiten der automatischen Indexierung muß doch festgestellt werden, daß für die englische Sprache die Algorithmen, Wortlisten und Softwaremodule schon recht weit sind. Für die - allerdings anders strukturierte - deutsche Sprache haben wir kaum Vergleichbares anzubieten. Es besteht die Gefahr, daß wir uns allein aus Sprachgründen vom informationswissenschaftlichen und -praktischen Fortschritt abkoppeln.

Für Carol Tenopir und Pamela Cahn ist nach ausgiebigen Retrievaltests an Freestyle und Target klar: "For now relevance retrieval is another weapon in the good searcher's arsenal". Für Online-Archive, die bislang Zurückhaltung bei natürlichsprachiger Retrievalsoftware üben, sei ein Umdenken empfohlen. Wir begrüßen "More like this"! ■

Wolfgang G. Stock

Auf dem ersten Blick:

FREESTYLE / MORE

Preis-Leistungsverhältnis	✪✪✪✪✪
Bedienungskomfort	✪✪✪✪
Qualität der Inhalte	✪✪✪

Weiterführende Informationen

http://www.lexis.nexis.com

LEXIS-NEXIS
Information Services GmbH
Lindenstr. 37
60325 Frankfurt
Tel.: 069/740534
Fax: 069/740638

Literaturhinweise

Susanne N. Bjoerner: The .WHERE and .WHY of FREESTYLE. - In: *Online 18 (1994), Nr. 2, 88 ff.*

Everett H. Brenner: *Beyond Boolean - New Approaches to Information Retrieval.* - Philadelphia, PA: The National Federation of Abstracting and Information Services, 1996.

Ross Evans: Beyond Boolean: Relevance Ranking, Natural Language and the New Search Paradigm. - In: *Proceedings of the Fifteenth National Online Meeting, New York, May 10-12, 1994. - Medford, NJ: Learned Information, 1994, 121 ff.*

Cary Griffith: LEXIS-NEXIS Goes Natural. - In: *Information Today 11 (1994), Nr. 1, 31+35.*

Sperry Krueger: Getting More out of NEXIS Using the MORE Search Command. - In: *Online 20 (1996), Nr. 1, 48 ff.*

Xin A. Lu; David J. Miller; John R. Wassum: *Phrase Recognition Method and Apparatus.* - US Pat. Nr. 5.819.260, erteilt am 6. Oktober 1998.

Joseph P. Mehrle: *Computer-Based System for Classifying Documents into a Hierarchy and Linking the Classifications to the Hierarchy.* - US Pat. Nr. 5.794.236, erteilt am 11. August 1998.

David J. Miller: *Advanced Freestyle Searching With Lexis-Nexis.* - Dayton: Lexis-Nexis, 1997.

Mick O'Leary: Freestyle Liberates Mead's Data. - In: *Searcher 2 (1994), Nr. 1, 45 ff.*

Lee Anne H. Paris; Helen R. Tibbo: Freestyle Vs. Boolean: A Comparision of Partial and Exact Match Retrieval Systems. - In: *Information Processing & Management 34 (1998), 175ff.*

Teresa Pritchard-Schoch: Comparing Natural Language Retrieval: WIN & FREESTYLE. - In: *Online 19 (1995), Nr. 4, 83 ff.*

Gerald Salton; Michael J. McGill: *Information Retrieval - Grundlegendes für Informationswissenschaftler.* - Hamburg [u.a.]: McGraw-Hill, 1987.

Carol Tenopir: Target, Freestyle, WIN ... Searching Takes on a New Look. - In: *Library Journal 119 (1994), Nr. 4, 34 f.*

Carol Tenopir; Pamela Cahn: TARGET & FREESTYLE: DIALOG and Mead Join the Relevance Ranks. - In: *Online 18 (1994), Nr. 3, 31 ff.*