Vorhersage der Suchwortwahl von professionellen Rechercheuren in Literaturdatenbanken durch assoziative Wortnetze

Reginald Ferber

Fachbereich 2
Postfach 1621
Universität –GH– Paderborn
D - 4790 Paderborn

Inhalt

Abstract

An associative lexical net is constructed to predict the selection of terms used in 94 professional, request based searches in a bibliographic data base. The weights of the net are calculated by a proportional learning rule, using the frequencies of co-occurence of terms in the free text fields of the 246,889 documents of the retrieval system PsycLIT (1989). To evaluate the results of the simulation the mean ranks of three classes of terms are calculated: Those terms that appear in both, the request and the query obtain a mean rank of 19 out of 872, those which appear in the query but not in the request, the new words, obtain a mean rank of 164, and those terms, that are not used in the query but appear in the request obtain a mean rank of 199.

Zusammenfassung

Mit einem assoziativen Wortnetz wird die Wortwahl bei 94 professionellen Recherchen, die aufgrund von schriftlichen Nutzeranfragen in einer Literaturdatenbank durchgeführt wurden, simuliert. Das Vokabular des Wortnetzes wurde auf der Basis der Anfragen und Recherchen konstruiert. Die Gewichte zwischen den Termen des Vokabulars wurden aus den Häufigkeiten des gemeinsamen Auftretens der Terme in den Freitextfeldern der Datenbank PsycLIT (1989) berechnet. Zur Bewertung der Simulationsergebnisse wurden die mittleren Rangplätze von drei Termklassen berechnet: Die Terme, die jeweils in der Anfrage und der zugehörigen Recherche auftauchen, erhielten bei 872 Termen einen mittleren Rang von 19, die, die in der Recherche auftauchen aber nicht in der Anfrage, erzielten einen mittleren Rang von 164, und die Terme, die nicht in die Recherche übernommen wurden, die also nur in der Anfrage auftauchen, lagen im Mittel auf Rang 199.

Diese Arbeit wurde vom Heinz-Nixdorf-Institut unterstützt

1. Wortwahl, Assoziationstheorie und Wortnetze

Zur Untersuchung der Interaktion zwischen Mensch und Maschine gehört insbesondere die Untersuchung des Verhaltens von Maschinenbenutzern in spezifischen Arbeitssituationen.

Im Folgenden wird ein assoziatives Modell zur Untersuchung der Wortwahl von professionellen Rechercheuren einer Literaturdatenbank bei der Generierung von Datenbanksuchfragen vorgestellt. Dabei wird zum einen die Auswahl von Worten aus natürlichsprachlichen schriftlichen Anfragen an die Datenbank untersucht, zum anderen die Wahl von Suchtermen, die nicht in der schriftlichen Anfrage vorkommen.

Die Wahl der verwendeten Suchterme wird mit der assoziativen Lerntheorie erklärt und durch ein daraus entwickeltes assoziatives Wortnetz mit gutem Erfolg simuliert. Die erfolgreiche Simulation der Wortwahl kann die Basis für eine interaktive Unterstützung von Rechercheuren oder eine Automatisierung der Suchtermgenerierung legen.

1.1. Assoziative Prozesse

Bei der Durchführung einer Recherche zu einer schriftlichen Anfrage müssen aus der Information der Anfrage Suchfragen in der Abfragesprache der Datenbank generiert werden. Dabei lassen sich zwei Arten von Prozessen unterscheiden: zum einen die Prozesse der Wortwahl, zum anderen die der Konstruktion der Suchfragen aus den gewählten Wörtern. Während es sich bei dem letzteren eher um einen regelgeleiteten Prozeß handelt, scheint die Wortwahl wesentlich durch assoziative Prozesse bestimmt.

Die Assoziationstheorie nimmt an, daß Dinge, die häufig zusammen wahrgenommen werden, im Gedächtnis derart miteinander verbunden sind, daß, wenn eines der beiden Dinge als Stimulus wahrgenommen oder erinnert wird, auch das andere erinnert wird. Auf die Sprache bezogen bedeutet das, daß Worte, die häufig zusammen auftreten, assoziativ verbunden sind, und umgekehrt, daß Worte, die assoziativ verbunden sind, in der Sprache häufig zusammen gebraucht werden. Zweifellos lassen sich viele Eigenschaften der Sprache, wie z. B. die syntaktische Struktur, mit diesem einfachen Ansatz nur schwer erklären, aber für die Vorhersage der Wortwahl scheint er geeignet zu sein.

Das assoziative Modell der Wortwahl läßt sich so zusammenfassen: Durch die Wahrnehmung von (fachspezifischer) Sprache werden im Laufe der Zeit (fachspezifische) Assoziationen zwischen den Wörtern der (Fach-) Sprache gelernt. Wird nun eine Anfrage gelesen, so werden die Wörter der Anfrage aktiviert. Diese Aktivierung setzt sich über die Assoziationen zum einen auf neue Wörter fort, die stark mit den aktivierten Wörtern der Anfrage verbunden sind, zum anderen stabilisieren sich die Aktivitäten von Wörtern, die stark untereinander verbunden sind, gegenseitig, während Wörter, die nur schwach mit den anderen aktivierten Wörtern verbunden sind, ihre Aktivierung verlieren. Als Suchterme werden schließlich die Wörter mit einer hohen Aktivierung verwendet.

1.2. Assoziative Wortnetze

Zur Modellierung dieses Ansatzes läßt sich ein sogenanntes assoziatives Wortnetz verwenden. Ein solches Netz besteht aus einer Menge von Knoten, denen eine reellwertige “Aktivierung” zugeordnet ist, und einer Menge von Verbindungen zwischen je zwei Knoten, denen ebenfalls je eine reelle Zahl als “Gewicht” zugeordnet ist. Den Knoten werden Wörter zugeordnet, die Gewichte zwischen zwei Knoten leiten sich aus den Assoziationen zwischen den Wörtern her. Ein assoziativer Prozeß wird als Ausbreitung von Aktivierungen auf dem Netz simuliert: Die Aktivitätswerte der Knoten, die dem Stimuluswort zugeordnet sind, werden erhöht und anschließend wird für jeden Knoten des Netzes ein neuer Aktivitätswert als Funktion der Verbindungsgewichte und der Aktivitäten der anderen Knoten berechnet. Dabei können die genauen Regeln der Aktivitätsausbreitung von Modell zu Modell verschieden sein.

In dem in dieser Untersuchung verwendeten linearen Modell besteht zwischen jedem Knotenpaar eine Verbindung, und der neue Wert eines Knotens ist die Summe der mit den jeweiligen Verbindungsgewichten multiplizierten Aktivitäten der anderen Knoten. Das entspricht der Multiplikation eines Vektors, der die Aktivierungen der Knoten enthält, mit einer quadratischen Matrix, die die Gewichte der Verbindungen zwischen den Knoten enthält.

1.3. Ziel der Untersuchung

Das hier vorgestellte Modell soll Vorhersagen über die Wortwahl von professionellen Rechercheuren ermöglichen und damit menschliches Verhalten modellieren. Ein erfolgreiches Modell könnte dazu dienen, Expertenwissen auch unerfahrenen Nutzern einer Datenbank zugänglich zu machen. Die Untersuchung unterscheidet sich von vielen Arbeiten zum Information Retrieval (z. B. [3], [4], [5]), dadurch, daß natürliche Sprache mit einem verhältnismäßig großen Vokabular verarbeitet wird und nicht Vektoren mit (möglichst linear unabhängigen) Ratings. Die dabei verwendeten Assoziationen werden aus großen maschinenlesbaren Textkorpora aus dem Fachgebiet der Datenbank gewonnen und sind folglich statistisch verhältnismäßig gut abgesichert.

2. Die Untersuchung

Im Folgenden wird die Untersuchung schrittweise dargestellt. Dabei werden auch einige theoretische Vorüberlegungen erst an den Stellen eingebracht, an denen sie inhaltlich zum Tragen kommen.

Die Simulation wurde auf einem PC weitgehend automatisiert durchgeführt. Der stärkste inhaltliche Eingriff ist sicherlich die Konstruktion der Knotenmenge. Ansonsten bewegte sich die Untersuchung aufgrund der großen zu verarbeitenden Datenmengen und der daraus resultierenden langen Rechenzeiten an der Grenze der Leistungsfähigkeit des gewählten Systems.

2.1. Das empirische Material

Grundlage für die Untersuchung waren 94 Literaturanfragen an die Zentralstelle für psychologische Information und Dokumentation in Trier, sowie die Protokolle der für die Anfragen durchgeführten Recherchen in den Datenbanken PsycINFO und PSYNDEX, sowie vereinzelt auch MEDLARS und SOCIAL SCISE.

Das Vokabular der Anfragen und Recherchen wurde als Knotenmenge des Wortnetzes verwendet. Die Gewichte wurden aufgrund des gemeinsamen Auftretens der Terme des Vokabulars in den Freitextfeldern der Dokumente der psychologischen Datenbank PsycLIT (1989) berechnet.

2.1.1. Die schriftlichen Anfragen und die Rechercheprotokolle

Die schriftlichen Anfragen und die zugehörigen Rechercheprotokolle stammen aus dem normalen Betrieb der Zentralstelle und entstanden unabhängig von dieser Untersuchung.

2.1.1.1. Der Anfragebogen

Für Literaturanfragen bei der Datenbank wird ein Formular benutzt, das für die Anfrage selbst zwei Felder vorsieht. Eines mit dem Titel: “Inhaltliche Beschreibung der Fragestellung in Form eines Arbeitstitels in deutscher, möglichst auch in englischer Sprache” und eines mit dem Titel: “ Suchstichworte möglichst aus der anglo-amerikanischen Fachsprache”. Im ersten Feld finden sich häufig natürlichsprachliche Texte, in denen die Nutzer ihre Fragestellung darlegen. Im zweiten Feld finden sich in der Regel Stichwortlisten, die teilweise aus ausgefallenen Wortkonstruktionen bestehen. Der gesamte Text aus diesen beiden Feldern wurde elektronisch erfaßt und für die Simulation verwendet. Im folgenden wird dieser Text als Nutzeranfrage oder kurz Anfrage bezeichnet. Er enthält im Durchschnitt $17,2$ Terme.

2.1.1.2. Die Rechercheprotokolle

Aufgrund der schriftlichen Anfragen an die Literaturdatenbank werden von den Rechercheurinnen und Rechercheuren der Zentralstelle Recherchen durchgeführt. Alle Rechercheure und Rechercheurinnen haben ein Diplom in Psychologie. Die Suchfragen, die im Laufe der Recherche verwendet wurden, wurden buchstabengetreu erfaßt. In ihnen wurden Trunkierungen, Feldspezifikationen und Content Codes (CC Nummern) verwendet. Die Kommandos der Abfragesprache wurden als solche gekennzeichnet und, ebenso wie die Content Codes, in der Auswertung nicht weiter berücksichtigt. (Mit Ausnahme des NOT Operators, siehe unten.) Die Recherchen enthielten im Durchschnitt $8,0$ Terme.

2.1.2. Das Textkorpus

Zur Bestimmung der Gewichte wurde die Literaturdatenbank PsycLIT (1989) verwendet. Diese Datenbank liegt auf CD-ROM vor und enthält die Einträge zu Zeitschriftenartikeln aus der Datenbank PsycINFO der Amerikanischen Psychologischen Gesellschaft. Die verwendete Ausgabe umfaßt 246 889 Dokumente, die zusammen ca. 310 Megabyte Text ausmachen.

Jedes Dokument gliedert sich in verschiedene Felder, die bibliographische Angaben, inhaltliche Klassifikationen, eine Zusammenfassung (AB) sowie eine stichwortartige Kurzzusammenfassung (KP) des Artikels enthalten. Für die Bestimmung der Gewichte wurden nur die Titel und die Zusammenfassungen der Dokumente verwendet, also nur solche Felder, die natürliche Sprache enthalten, und nicht solche mit Klassifikationen. Die Titel sind in der Originalsprache und gegebenenfalls auch in englischer Übersetzung angegeben, die Zusammenfassungen sind englisch.

Inhaltliche Beschreibung der Fragestellung in Form eines Arbeitstitels in deutscher, möglichst auch in englischer Sprache: `Einfluß von Geschlecht stereotypen auf sexuelles Verhalten. Influence of sex role stereotypes on sexual behavior`
Suchstichworte möglichst aus der anglo-amerikanischen Fachsprache: `1.) Sex role stereotype 2.) Androgyny`

```
PSYNDEX
1 177 Find CT All (Sex Role Att$;Feminism;Feminity;Masculinity)
2 11 Find ALL Androgyn$/PQ
3 734 Find CT D Psychosexual Behavior
4 428 Find 3 Not CT=Sex Roles
5 17 Find (1;2) And 4
```

Beispiel 1: Eine Nutzeranfrage und die zugehörige Recherche. Das Wort "Geschlechtstereotypen" wurde wie in 2.2.1.2 beschrieben in die Terme "Geschlecht" und "Stereotypen" aufgespalten.

2.2. Durchführung der Simulation

Zur Durchführung der Simulation wurde auf der Grundlage des empirischen Materials ein assoziatives Wortnetz konstruiert. Dazu wurde zunächst eine Menge von Knoten benötigt.

2.2.1. Das Vokabular

Die Knotenmenge für das assoziative Wortnetz wurde aus dem Vokabular der Nutzeranfragen und der Rechercheprotokolle konstruiert. Aus den folgenden Gründen wurden dabei jeweils mehrere Wörter in einem Knoten zusammengefaßt:

2.2.1.1. Größe des Vokabulars

Bei dem verwendeten Modell des Wortnetzes bestehen zwischen allen Knoten jeweils gewichtete Verbindungen. Die Anzahl der zu verarbeitenden Gewichte wächst folglich mit dem Quadrat der Knotenanzahl. Um die Rechenzeiten in Grenzen zu halten, sollte die Anzahl der Knoten deshalb nicht zu groß werden. Zum anderen werden die Gewichte zwischen den Knoten aufgrund des Auftretens der zugeordneten Wörter in einer begrenzten Textmenge geschätzt. Für Knoten mit seltenen Wörtern ist eine solche Schätzung vergleichsweise ungenau. Deshalb sollten die Wörter eines Knotens eine gewisse Mindesthäufigkeit haben, wenn der Knoten in die Simulation einbezogen werden soll.

2.2.1.2. Mehrsprachigkeit

Die Benutzeranfragen waren überwiegend zweisprachig deutsch und englisch formuliert, die Suchfragen überwiegend englisch formuliert. Die Zusammenfassungen in den Dokumenten der Datenbank waren dagegen alle englisch. Deshalb wurden deutsche Wörter und ihre englischen

Übersetzungen zu einem Knoten zusammengefaßt. Gab es zu einem deutschen Wort keine englische Übersetzung in dem Wortschatz der Nutzeranfragen und Rechercheprotokolle, so wurde sie hinzugefügt. Bestand die englische Übersetzung eines deutschen Wortes aus mehreren Wörtern, so wurde das deutsche Wort in den Protokollen entsprechend den englischen Wortgrenzen getrennt. Die einzelnen Wortteile wurden den entsprechenden Knoten zugeordnet. (So wurde beispielsweise “Sozialverhalten” in “sozial” und “Verhalten” aufgetrennt.)

2.2.1.3. Wortformen und Lemmatisierung

Die Nutzeranfragen wurden wörtlich erfaßt, d. h. die Wörter traten in den entsprechenden grammatikalischen Formen auf. Da zum einen die spezielle Form eines Wortes nur bedingt Einfluß auf die mit ihm verbundenen Assoziationen haben sollte, und zum anderen die Erfassung jeder einzelnen Wortform in einem separaten Knoten die Anzahl der Knoten sehr vergrößern würde, wurden die verschiedenen Formen eines Wortes ebenfalls zu einem Knoten zusammengefaßt.

2.2.1.4. Die Konstruktion des Vokabulars

Um die obigen Gesichtspunkte zu berücksichtigen, wurden die Knoten für das Wortnetz in der folgenden Weise konstruiert:

Zunächst wurden aus den Abschriften alle Zeichenfolgen isoliert, die nur Buchstaben enthielten. Daraus wurden Buchstabenfolgen entfernt, die keine englischen oder deutschen Wörter bezeichneten, also z. B. die Feldbezeichner der Datenbank oder Namen. Nicht entfernt wurden Zeichenketten, die als Trunkierungen von Wörtern oder als Wörter mit Tippfehlern erkennbar waren. Die übrigbleibenden Wörter bildeten das Grundvokabular.

In einem nächsten Schritt wurden verschiedene Formen eines Wortstammes, Wörter mit Tippfehlern und die jeweiligen Übersetzungen zu einem Knoten zusammengefaßt. Dabei wurde zu deutschen Wörtern zunächst eine englische Übersetzung innerhalb des Grundvokabulars gesucht. Falls diese nicht vorhanden war, wurde sie nach Collins German Dictionary [6] hinzugefügt. Trunkierte Wörter wurden einem Knoten zugefügt, der eine passende Expansion enthielt. Es trat nicht auf, daß zu einer Trunkierung keine Expansion im Grundvokabular vorhanden war. Bei englischen Wörtern, zu denen keine deutsche Übersetzung im Grundvokabular vorlag, wurde keine Übersetzung hinzugefügt, da das Korpus ja ohnehin kaum deutsche Wörter enthielt (allenfalls in den Titeln der Artikel). Schließlich wurden zu allen englischen Substantiven die Singular- bzw. Pluralformen hinzugefügt, soweit sie existierten.

Ein Knoten des Vokabulars enthielt also mindestens ein englisches Wort. Er konnte verschiedene Formen eines Wortstammes sowie die Übersetzung des Wortes und deren Formen enthalten. (Die Knoten, die in Beispiel 1 auftreten, sind beispielsweise in Tabelle 2 angegeben.) Nach Abschluß der Konstruktion bildete das Vokabular 947 Knoten.

Schließlich wurde für jeden Knoten bestimmt, in wievielen Dokumenten des Textkorpus er vorkam. Dabei wurde immer dann vom Vorkommen eines Knotens in einem Dokument ausgegangen, wenn eines der Worte des Knotens in einem der Freitextfelder des Dokuments, also im Titel, in der Zusammenfassung oder in den Stichworten vorkam. 20 Knoten kamen in keinem Dokument vor. 55 Knoten kamen nur in weniger als 40 Dokumenten vor. Die verbleibenden 872 Knoten wurden schließlich zur Konstruktion des Wortnetzes verwendet.

2.2.2. Die Gewichte

Aus der Lernfomel des Assoziationsgesetzes läßt sich eine Formel zur Berechnung der Gewichte aus den Auftretenshäufigkeiten ableiten: Nimmt man an, daß sich die Assoziation $a_{i,j}(t)$ von einem Wort i zu einem Wort j beim Wahrnehmen von Text immer dann erhöht, wenn i und j zusammen auftreten, und daß sich $a_{i,j}(t)$ andererseits verringert, wenn das Wort i ohne das Wort j auftritt, so läßt sich das, bei Assoziationstärken zwischen 0 und 1, folgendermaßen zu einer proportionalen Lernregel formalisieren:

$$a_{i,j}(t+1) = \begin{cases} a_{i,j}(t) + \alpha \cdot (1 - a_{i,j}(t)) & \text{falls } (i\&j) \\ (1-\alpha) \cdot a_{i,j}(t) & \text{falls } (i\&\neg j) \end{cases} \quad (1)$$

Dabei ist die Lernrate α ein Parameter, der die Stärke der Veränderung steuert. $(i\&j)$ stellt das Ereignis dar, daß i und j zusammen auftreten, $(i\&\neg j)$ das, daß i ohne j auftritt.

Falls die Reihenfolge der Ereignisse $(i\&j)$ und $(i\&\neg j)$ zufällig ist, ist $a_{i,j}(t)$ für $t \to \infty$ eine Schätzung der bedingten Wahrscheinlichkeit $p(j \mid i)$ des Auftretens von j unter der Bedingung i (Vgl. z. B. [2]). Diese Wahrscheinlichkeit läßt sich aber auch durch die relative Häufigkeit

$$g_{i,j} = \frac{H(i\&j)}{H(i)} \tag{2}$$

abschätzen. Dabei bezeichnet $H(x)$ die Häufigkeit des Ereignisses x, d. h. im Fall der Simulation beispielsweise für $x = (i\&j)$, in wievielen Dokumenten der Datenbank i und j zusammen vorkommen.

Die Formel (2) berücksichtigt allerdings nicht die Häufigkeit des Wortes j. Die bedingte Wahrscheinlichkeit eines häufigen Wortes unter der Bedingung eines seltenen ist aber groß, was dazu führen kann, daß die mit Formel (2) berechneten Gewichte von seltenen zu häufigen Knoten groß sind und dadurch häufige (und damit eher unspezifische) Knoten leichter aktiviert werden als seltene. Dem kann dadurch entgegengewirkt werden, daß zusätzlich durch die Häufigkeit $H(j)$ von j dividiert wird. So erhält man

$$g_{i,j} = \frac{H(i\&j)}{H(i) \cdot H(j)} \tag{3}$$

Diese Formel erhält man, bis auf einen konstanten Faktor, auch, wenn man den Quotienten aus der Wahrscheinlichkeit $p(i\&j)$ für das gemeinsame Auftreten von i und j und dem Produkt $p(i) \cdot p(j)$, der Auftretenswahrscheinlichkeit von i zusammen mit j im Falle ihrer Unabhängigkeit, bildet und dann wieder relative Häufigkeiten für die Wahrscheinlichkeiten einsetzt. Im Fall der Unabhängigkeit von i und j ist der Quotient 1. Will man vermeiden, daß in diesem Fall Aktivierung von i nach j weitergeleitet wird, kann man den Wert bei Unabhängigkeit von der Formel abziehen und erhält schließlich

$$g_{i,j} = \frac{H(i\&j) \cdot A}{H(i) \cdot H(j)} - 1 \tag{4}$$

wobei A die Anzahl aller Dokumente ist. Der Wertebereich dieser Formel liegt zwischen -1 und $A-1$, wobei kleine Schwankungen in der Schätzung der relativen Häufigkeiten der Einzelknoten bei kleinen Werten starke Schwankungen der Werte der Gewichte bewirken. Um diesen Effekt zu dämpfen, kann eine nichtlineare monoton wachsende hyperbolische Transformation auf den Bereich $[-1, 1]$ durchgeführt werden.

Die verschiedenen Formeln und Transformationen wurden als freie Parameter in die Simulation eingeführt. Die nichtlineare Transformation enthielt als weitere freie Parameter die Steigung im Ursprung und einen Faktor auf die negativen Gewichte.

Diese Formeln können auch für $i = j$ verwendet werden. Es ergibt sich dann für (3)

$$g_{i,i} = \frac{H(i\&i)}{H(i) \cdot H(i)} = \frac{H(i)}{H(i) \cdot H(i)} = \frac{1}{H(i)} \tag{5}$$

und für (4)

$$g_{i,i} = \frac{H(i\&i) \cdot A}{H(i) \cdot H(i)} - 1 = \frac{A}{H(i)} - 1 \tag{6}$$

Verwendet man bei der Berechnung der neuen Aktivitäten der Knoten für $i = j$ den selben Algorithmus wie für $i \neq j$, so wird die Aktivität eines Knotens mit dem Faktor $g_{i,i}$ multipliziert übernommen. Für $0 \leq g_{i,i} \leq 1$ stellt das ein langsames Abklingen der Aktivierung des Knotens dar.

Es wurde ein freier Parameter eingeführt, der die Behandlung der Gewichte $g_{i,i}$ steuert.

2.2.3. Die Iteration des Wortnetzes

Mit der Bestimmung der Knotenmenge und der Gewichte ist die Konstruktion des assoziativen Wortnetzes im Prinzip abgeschlossen. Die Simulation des Wortwahlprozesses geht nun so vor sich, daß die Knoten der Wörter der Nutzeranfrage aktiviert werden und dieses Aktivierungsmuster über mehrere Zyklen auf dem Netz weiterentwickelt wird. Da es sich bei dem verwendeten Wortnetz aber um ein lineares System handelt, lassen sich einige Vorhersagen über sein Verhalten machen, die eine gewisse Normierung des Netzes nahelegen.

2.2.3.1. Das Wortnetz als lineares System

Wie schon in 1.2 erwähnt, lassen sich die verwendeten Wortnetze als Multiplikation

$$y = G \cdot x \tag{7}$$

eines Aktivitätsvektors x mit einer quadratischen Matrix G, die die Gewichte enthält, also als mehrdimensionale lineare Abbildung verstehen. Der Akitvitätsvektor ist von der Länge der Anzahl der Knoten und enthält als k-ten Eintrag die Aktivität des k-ten Knotens. Zu Beginn der Simulation wird ein Inputvektor b gebildet, der an den Stellen der Knoten der Stimuluswörter einen positiven Eintrag enthält und sonst nur Nullen. Eine n–malige Iteration des Prozesses führt zur folgenden Formel

$$x^n = G \cdot x^{n-1} = G\big(G \cdot x^{n-2}\big) = G^n \cdot b \tag{8}$$

Wird davon ausgegangen, daß der Stimulus dauernd präsent ist, so wird der Inputvektor b bei jedem Iterationsschritt addiert und man erhält:

$$x^n = G \cdot \big(x^{n-1} + b\big) = G \cdot \big(G \cdot \big(x^{n-2} + b\big) + b\big) = G^n \cdot b + G^{n-1} \cdot b + ... + G \cdot b \tag{9}$$

2.2.3.2. Die Normierung der Gewichte

Unter (8) konvergiert x^n für $n \to \infty$ im allgemeinen gegen den größten Eigenvektor von G und ist daher weitgehend unabhängig vom tatsächlichen Input. Unter (9) hängt die Entwicklung von x^n vom Spektralradius der Matrix G ab. Ist er größer als 1, so dominiert der Term $G^n \cdot b$ die Entwicklung von x^n, ist er kleiner als 1, dominiert $G \cdot b$ die Entwicklung. Deshalb wurde für die Simulation (9) verwendet und der Spektralradius der Matrix auf 1 normiert, indem alle Gewichte durch den Spektralradius der ursprünglichen Gewichtsmatrix dividiert wurden. Dadurch sollte sich in (9) der Einfluß der Summanden ungefähr die Waage halten. Zusätzlich wurde als freier Parameter ein Faktor auf die Matrix eingeführt, der die Gewichtung der Terme in (9) steuerbar macht.

2.2.3.3. Durchführung der Simulation

Die Simulation der Wortwahl für eine Nutzeranfrage wurde nun folgendermaßen durchgeführt: Zuerst wurden die Aktivitäten aller Knoten auf 0 gesetzt. Dann wurde die Nutzeranfrage automatisch eingelesen. Dabei wurde jedesmal, wenn ein Term des Vokabulars gefunden wurde, die Akitivität des zugehörigen Knotens um 1 erhöht. Der so gebildete Inputvektor wurde gespeichert. Dann wurde ein Iterationszyklus des Netzes ausgeführt. D. h. es wurden auf der Grundlage des Inputvektors neue Aktivitäten für alle Knoten berechnet. Der neue Aktivitätsvektor wurde gespeichert, um später ausgewertet zu werden. Daraufhin wurde zu dem Aktivitätsvektor der Inputvektor addiert und ein weiterer Iterationszyklus des Netzes berechnet.

2.3. Ergebnisse

Zur Bewertung der Ergebnisse der Simulation wurden die Terme der Knoten mit hohen Aktivitäten mit den Termen verglichen, die in den Recherchen vorkamen. Dazu wurde folgendes Auswertungsschema verwendet:

2.3.1. Auswertungsschema

Zunächst wurden die Terme des Vokabulars für jede Nutzeranfrage in vier Klassen eingeteilt:

2.3.1.1. Übernommene oder IN-IN Terme

Eine Klasse bildeten die Terme, die sowohl in der Nutzeranfrage als auch in der Recherche vorkamen, die also aus der Nutzeranfrage in die Suchfragen übernommen wurden. Sie sollten eine hohe Aktivität und damit einen niedrigen Rangplatz haben.

2.3.1.2. Neue oder OUT-IN Terme

Die nächste Klasse bildeten die Terme, die nur in den Suchfragen, nicht aber in der Nutzeranfrage vorkamen, die der Rechercheur oder die Rechercheurin also neu gewählt hat. Sie sollten ebenfalls eine hohe Aktivität und damit einen niedrigen Rangplatz haben.

2.3.1.3. Nicht übernommene oder IN-OUT Terme

Die dritte Klasse sind die Terme, die nicht in die Suchfragen übernommen wurden, die also nur in der Nutzeranfrage auftauchten. Sie sollten geringere Aktivitäten und damit höhere Rangplätze haben.

2.3.1.4. Nicht auftauchende oder OUT-OUT Terme

Schließlich bleibt noch die große Mehrzahl aller übrigen Terme, die weder in der Nutzeranfrage noch in den Suchfragen auftauchen. Ihre Aktivität sollte gering sein.

Zur Auswertung der Simulation wurden die Knoten anhand der während der Iteration des Netzes gespeicherten Aktivitätsvektoren für jeden Iterationszyklus nach ihrer Aktivität in eine Rangreihe gebracht und jeweils die mittleren Rangplätze der Terme aus den vier Termklassen berechnet. Bei der Mittelwertbildung wurden die Terme in der Nutzeranfrage mit ihrer Vielfachheit gezählt.

Zur Bewertung mehrerer Simulationen wurden die mittleren Rangplätze der Termklassenelemente über die verschiedenen Simulationen berechnet.

Die Bewertung wurde noch weiter differenziert, indem Terme, die mit dem NOT-Operator verwendet wurden, getrennt ausgewertet wurden. Dazu wurden sie in die oben definierten Klassen eingeteilt. Es kam allerdings häufig vor, daß im Verlauf der Recherche ein Term sowohl mit als auch ohne NOT-Operator verwendet, bzw. daß mit Referenznummern auf die Ergebnisse früherer Suchfragen zurückgegriffen wurde und diese mit NOT in die aktuelle Suchfrage einbezogen wurden. In diesen Fällen tauchten die Terme in beiden Klassensystemen auf.

2.3.2. Auswertung

Bei der Konstruktion des Wortnetzes waren diverse Parameter eingeführt worden. Um sie zu optimieren wurden zunächst 47 Protokolle zufällig ausgewählt und an diesen die Parameter nach verschiedenen Kriterien optimiert (Einzelheiten siehe [1]). Dann wurden zur Überprüfung die anderen 47 Protokolle mit den Parametern simuliert, mit denen bei der ersten Hälfte der Protokolle die besten Ergebnisse erzielt worden waren.

Wenn das Modell eine allgemeine Gültigkeit haben soll, müssen die Ergebnisse für beide Hälften etwa gleich gut sein. Die Tabellen 3 und 4 zeigen, daß dies weitgehend der Fall ist.

2.4. Diskussion

Die Gesamtergebnisse der Simulation für alle Terme sind in den Tabellen 3 und 4 und die für die Terme, die mit dem NOT-Operator verwendet wurden, in den Tabellen 5 und 6 angegeben.

Rang	Klasse	Aktivierung	Term
1	in-out	0.009727	STEREOTYPEN STEREOTYPES STEREOTYPE
2	in-in	0.009691	ANDROGYN ANDROGYNY
3	out-in	0.008722	MASCULINITY
4	in-out	0.007178	GENDER GESCHLECHT
5	in-in	0.007116	SEX SEXUAL SEXUALITAT SEXUELLE SEXUELLES
6	out-in	0.005652	FEMINIS FEMINISM
7		0.004986	HOMOSEXUALITY
8		0.004283	MASTURBATION
9	in-in	0.003922	ROLE ROLES ROLLE ROLLEN
10		0.00378	INTERCOURSE intercourses
11		0.003707	LIBERAL
12		0.003591	MAN MEN
13		0.00356	FRAU FRAUEN WOMAN WOMEN
14	in-out	0.003453	EINFLUA EINFLUSSEN INFLUENCE INFLUENCING INFLUENCES
15		0.003453	ORGASM
16	out-in	0.00328	PSYCHOSEXUAL
17		0.003242	OCCUPATION
18		0.003032	IDENTITY
19		0.002815	PARTNER PARTNERN PARTNERS PARTNERSCHAFT PARTNERSCHAFTLICH
20		0.002813	FEMALE FEMALES WEIBLICHEN
56	in-in	0.001462	BEHAVIOR BEHAVIORAL BEHAVIORS VERHALTEN VERHALTENS
195	in-out	0.000579	AUF ON
225	in-out	0.000448	VON FROM VOM
439	in-out	0.000008	OF AUS

Tabelle 2: Ergebnisse der (erfolgreichen) Simulation zur Nutzeranfrage aus Beispiel 1. Die Knoten wurden nach ihren Aktivierungen sortiert. Die ersten 20 Rangplätze sind vollständig angegeben, danach nur noch die Rangplätze von Termen, die in der Anfrage oder der Recherche vorkommen. (Zur Bezeichnung der Klassen siehe 2.3.1.)

Zyk.	in-in	out-in	in-out	alle Query
1	18.5	155.6	194.4	50.7
2	24.5	156.0	216.0	55.4
3	29.0	158.6	227.6	59.5
4	32.2	161.2	234.7	62.5
5	34.4	162.9	239.3	64.6

Tabelle 3: Die mittleren Rangplätze der Termklassen für die Protokolle, an denen die Parameter optimiert wurden. Unter "alle Query" sind die mittleren Rangplätze der Terme aus der Query, also aus der Vereinigung der in-in und der out-in Klasse, aufgeführt. Insgesamt werden die Rangplätze im Laufe der Iteration höher, allerdings wachsen sie für die in-out Terme schneller als für die out-in Terme.

2.4.1. Beurteilung der Ergebnisse

Es zeigt sich, daß die Ergebnisse zwar noch ein gutes Stück von der Wortwahl der Rechercheu-

Zyk.	in-in	out-in	in-out	alle Query
1	18.9	172.2	203.7	55.1
2	22.9	180.3	227.8	60.1
3	25.7	188.7	241.1	64.2
4	27.7	194.7	249.7	67.1
5	29.0	198.3	255.2	69.0

Tabelle 4: Die mittleren Rangplätze der Termklassen für die Protokolle, an denen die Parameter nicht optimiert wurden, die für das System also völlig neu waren. Es zeigt sich eine gute Übereinstimmung mit den Werten aus Tabelle 3.

Zyk.	in-in	out-in	in-out	alle Query
1	9.8	230.5	143.0	109.7
2	12.2	218.8	160.1	105.8
3	14.3	212.3	169.6	104.0
4	14.9	206.5	175.6	101.6
5	15.3	202.1	179.6	99.9

Tabelle 5: Die mittleren Rangplätze der Termklassen der Terme, die mit NOT verwendet wurden, für die Protokolle, an denen die Parameter optimiert wurden.

rinnen und Rechercheure entfernt sind, daß sie aber andererseits gute statistische Vorhersagen darstellen.

Bei ihrer Beurteilung sind mehrere Faktoren zu berücksichtigen:

– Die Anzahl der Terme des Vokabulars beträgt 872. Für ein zufälliges Ergebnis läge folglich der erwartete mittlere Rangplatz bei 436.
– Bei der Optimierung wurden die Parameter so gewählt, daß der mittlere Rangplatz der out-in Terme niedriger war als der der in-out Terme. Optimiert man die Parameter auf ein möglichst gutes Ergebnis der in-in Terme, so kann man einen Wert von $6,2$ für die Protokolle, mit denen optimiert wurde, und einen Wert von $7,5$ für die übrigen Protokolle erreichen.
– Beobachtet man die Entwicklung der Werte über die Zyklen, so werden die Terme der Recherchen zwar insgesamt schlechter, aber andererseits vergrößert sich der Abstand zwischen den out-in Termen und den in-out Termen im Laufe der Entwicklung.
– Die Simulation wird aufgrund der Terme der Recherchen beurteilt. Da die Daten aus dem normalen Betrieb der Literaturdatenbank stammen, liegt pro Anfrage nur eine Recherche vor. Es konnte deshalb nicht untersucht werden, wie stark sich möglicherweise die Wortwahl verschiedener Rechercheure und Rechercheurinnen zu ein und der selben Nutzeranfrage unterscheiden würden.

Bei den Termen, die mit dem NOT-Operator verwendet wurden, ist das Bild nicht so einheitlich wie bei dem Gesamtergebnis. Das liegt vermutlich an der geringen Zahl von 36 Termen bei den Protokollen für die Optimierung und 39 bei den übrigen Protokollen. Trotzdem läßt sich erkennen, daß während die in-in Terme besser plaziert sind als im Gesamtergebnis, die out-in Terme deutlich schlechter sind. Das könnte dadurch erklärt werden, daß die Prozesse, bei denen in den Recherchen versucht wird, durch Terme mit NOT-Operator bestimmte Dokumentmengen auszuschließen, eher regelgeleitete Prozesse sind.

2.4.2. Schwächen des Systems und mögliche Weiterentwicklungen

Das vorgestellte Modell weist in vielen Punkten noch Schwachstellen auf, die durch Weiterentwicklungen verbessert werden können.

Zyk.	in-in	out-in	in-out	alle Query
1	5.2	268.3	151.0	121.7
2	6.5	283.0	169.3	129.0
3	7.7	299.5	179.7	136.9
4	8.6	311.3	186.4	142.6
5	9.8	317.7	190.7	146.2

Tabelle 6: Die mittleren Rangplätze der Termklassen der Terme, die mit NOT verwendet wurden, für die Protokolle, an denen die Parameter nicht optimiert wurden.

2.4.2.1. Vokabular

Die Konstruktion des Vokabulars ist der Teil der Simulation, bei dem am meisten "von Hand" gearbeitet wurde. Dabei mußten an vielen Stellen Entscheidungen durch die Bearbeiter getroffen werden, die algorithmisch (noch) nicht zu lösen sind.

Insbesondere lassen sich folgende Punkte anführen:

– Das Vokabular wurde aus den zu simulierenden Protokollen gewonnen. Dadurch sind die dort vorkommenden Worte zwar weitgehend erfaßt, aber es fehlen auch viele häufige Worte (z. B. "Aggression"). Eine Vergrößerung des Vokabulars wirft aber beim momentan verwendeten vollständig vernetzten Modell technische Schwierigkeiten auf.
– Die Datenbank erlaubt zusammengesetze Suchterme. Das führt dazu, daß in den Suchfragen Wörter auftauchen, die für sich allein genommen wenig spezifisch sind, wie z. B. die Wörter "behavior" oder "role" in Beispiel 1. Diese Wörter werden von dem gegenwärtigen System als einzelne Terme behandelt und sind daher in-in terme, auch wenn sie einzeln wenig spezifisch sind. Bei einem System, das zusammengesetzte Terme zuläßt, sind allerdings diverse Probleme zu lösen, so z. B. welche Wortketten als Terme zugelassen werden, oder wie die einzelnen Wörter in den zusammengesetzten Termen behandelt werden sollen. Erste Untersuchungen mit einem solchen System zeigen keine wesentlichen Verbesserungen der Ergebnisse.
– Die Datenbank erlaubt die Verwendung von Trunkierungen. Die Trunkierungen wurden durch den Bearbeiter zu Knoten hinzugefügt, die passend erschienen. So wurde z. B. die Trunkierung "stud$" dem Knoten "stud, student, students, schüler" zugeordnet. In der Recherche war sie aber in dem Befehl "find all follow up stud$" verwendet worden.
– Bei der Übersetzung vom Deutschen ins Englische kann es zu Mehrdeutigkeiten kommen. So wurde das Wort "war" dem Knoten "war, was" zugeordnet, was bei der Nutzeranfrage "...following the second World War..." sicherlich zu einer falschen Zuordnung führt. Durch Tippfehler können ähnliche Effekte hervorgerufen werden.

2.4.2.2. Evaluierung

Zur Bewertung der Simulationsergebnisse wurden in der gegenwärtigen Untersuchung lediglich die Terme aus einer einzigen Recherche verwendet. Es wäre sicherlich sinnvoll, andere Evaluierungsmethoden zu entwickeln. So könnten zum Beispiel die vom System gewählten Terme Rechercheurinnen und Rechercheuren zur Einschätzung vorgelegt werden, ähnlich wie es bei der Evaluierung von Rechercheergebnissen getan wird.

Literatur

[1] R. Ferber, M. Wettler, and R. Rapp. An associative model of word selection in the generation of search queries. *Submitted for publication*, 1993.

[2] K. Foppa. *Lernen, Gedächtnis, Verhalten*. Köln, Berlin: Kiepenheuer & Witsch, 1965.

[3] V. E. Giuliano and P. E. Jones. Linear associative information retrieval. In P. W. Howerton and D. C. Weeks, editors, *Vistas in Information Handling*, volume 1, chapter 2, pages 30–54. Spartan Books, Washington D. C., Washington, D.C., 1963.

[4] W. P. Jones and G. W. Furnas. Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science*, 38(6):420–442, 1987.

[5] G. Salton and C. Buckley. On the use of spreading activation methods in automatic information retrieval. In *Proceedings of the 11th Annual International Conference on Research and Development in Information Retrieval*, pages 147–160. ACM, 1988.

[6] P. Terrell, V. Calderwood-Schnorr, W. V. A. Morris, and R. Breitsprecher, editors. *Collins German Dictionary*. Collins, London & Glasgow, 1984.