Uwe Quasthoff & Christian Wolff

Effizientes Dokumentenclustering durch niederfrequente Terme

Universität Leipzig
Institut für Informatik
04109 Leipzig
{quasthoff, wolff}@informatik.uni-leipzig.de

Abstract

In diesem Papier wird ein statistisches IR-Verfahren vorgestellt, mit dem sich ähnliche Dokumente in umfangreichen Kollektionen effektiv ermitteln lassen. Das Verfahren verwendet als Deskriptoren ausschließlich niederfrequente Terme. Die zur Dokumentbeschreibung benutzten Dokumentvektoren sind schwach besetzt (*sparse vectors*) und erlauben effiziente Berechnungsverfahren. Die Ermittlung geeigneter Deskriptoren zieht als zusätzliche Informationsquelle eine umfangreiche Datenbank mit Frequenzinformation heran.

1 Einleitung

Für *vector space*-basierte Anwendungen im Information Retrieval, beim Clustering von Dokumenten und der automatischen Hypertextgenerierung wird die dokumentenbezogene Information i.d.R. ausschließlich auf der Basis der Dokumentenkollektion selbst gewonnen (Vgl. SALTON & MCGILL 1983:59ff, KORFHAGE 1996:82ff, SALTON et al. 1996:52). Der Einsatz von semantischem Wissen in Form der Zuweisung niederfrequenter Terme zu Thesaurus(ober)klassen (vgl. SALTON 1989:308) dient üblicherweise eher der Gruppenbildung und Ergänzung als zu spezifisch erachteter niederfrequenter Terme. Der hier verfolgte Ansatz geht dagegen von folgenden Annahmen aus:

- Die Signifikanz von Deskriptoren läßt sich präziser bestimmen, wenn kollektions- und damit auch domänen- und textsortenunabhängig Information über die Auftretenswahrscheinlichkeit von Termen in beliebigem Text vorliegt.
- Der Vergleich der *document frequency* mit der *universal frequency* des Terms in der Gesamtheit aller Texte innerhalb wie außerhalb der Kollektion ist eine gute Basis für die Termselektion.
- Universelle Frequenzinformation ist erst mittelfristig mit hinreichender Güte zu erreichen, insbesondere ist anzunehmen, daß stark fachzentrierte Kollektionen momentan noch nicht so stark separiert werden wie Kollektion mit breiter Sachgebietsstreuung (enzyklopädisches Wissen, Zeitungskollektionen).

2 Lexikalische Hintergrundinformation und Termrelevanz

Zur Erzeugung der Dokumentenbeschreibung wird die im Projekt *Deutscher Wortschatz* aufgebaute Datenbank (vgl. QUASTHOFF 1998)[1] verwendet, ein Vollformenlexikon des Deutschen mit mehr als 3.000.000 Einträgen. Die Datenbank enthält u.a. grammatikalische Information sowie Frequenzangaben. Zusätzlich stehen in vielen Fällen Verweise von flektierten Formen auf die entsprechende Grundform zur Verfügung. Die Frequenzdaten wurden aus einem Korpus von ca. 250 Millionen laufenden Wörtern ermittelt, der

[1] Das Projekt *Deutscher Wortschatz* wurde im Rahmen des DFG-Projektes LAPT & DA (*dynamische Aktivierung domänenspezifischer Teillexika)* gefördert.

den ganzen Bereich derzeit elektronisch verfügbarer Dokumente abzudecken versucht. Damit läßt sich die Auftretenswahrscheinlichkeit eines Terms in Bezug auf den Gesamtkorpus mit der tatsächlichen Auftretenshäufigkeit im konkreten Dokument vergleichen und als Relevanzinformation nutzen.

Die Verwendung eines sehr großen Korpus zur Ermittlung der Frequenzdaten niederfrequenter Wörter scheint notwendig, um einerseits hinreichende Sicherheit bei der relativen Häufigkeit zu erlangen und andererseits große Veränderung der Frequenzdaten durch hinzukommende Dokumente zu vermeiden.

Für die Termrelevanz eines Deskriptors in der Kollektion wurde folgende logarithmische Skala gewählt: Als Relevanzmaß dient der duale Logarithmus des Quotienten von tatsächlicher Auftretenshäufigkeit eines Terms in einem Dokument und dem aus der Datenbank ermittelten Erwartungswert für die Auftretenshäufigkeit entsprechend der Dokumentlänge. Als *Relevanzklasse* eines Terms in einem Dokument soll die (ebenfalls logarithmisch gemessene) Überschreitung des Erwartungswertes bezeichnet werden.
Als unterer Schwellenwert für die Annahme der Relevanz eines Terms hinsichtlich seiner Eignung als Deskriptor wurde eine Relevanzklasse von mindestens 10 (also bezogen auf die Termfrequenzen mindestens um einen Faktor 10^3 höher als erwartet) gewählt. Dieser Schwellenwert ergab sich aus einer ersten Bewertung von Clusteringergebnissen und wurde so gewählt, um einerseits die Anzahl der relevanten Terme gering zu halten, aber andererseits noch sinnvolle Ähnlichkeitsaussagen machen zu können.

3 Die Testkollektion

Für das Clustern wurde eine Testkollektion zusammengestellt, die aus ca. 60.000 Dokumenten besteht, wobei jeweils zur Hälfte

- Zeitungsartikel der *Tageszeitung* aus dem Zeitraum 9/86 bis 12/87 [TAZ] und
- Lexikonartikel aus *Microsoft Encarta 98* [ENC]

verwendet wurden. Die mittlere Artikellänge beträgt 1.82 bzw. 1.62 KB, die zulässige maximale Länge wurde auf 30KB beschränkt.
Durch die gemischte Testkollektion ist gewährleistet, daß die semantische Struktur der Artikel äußerst unterschiedlich ist: Ein typischer Lexikonartikel mittlerer Länge behandelt einen Fachbegriff und nennt im Text relevante Begriffe des gleichen Fachgebiets. Anders bei vielen Zeitungsartikeln: Hier wird oft ein Zusammenhang zwischen zwei Themen hergestellt (z.B. ein Ereignis und dadurch hervorgerufene Reaktionen, die Meinung einer Person/Institution zu einem Problem etc.). Demzufolge sind in einem solchen Dokument Terme aus mindestens zwei Gebieten relevant, das entsprechende Dokument weist also Ähnlichkeiten zu Dokumenten aus mehreren Klassen auf.

4 Dokumentenbeschreibung

Zur Dokumentenbeschreibung werden die relevanten Terme zusammen mit ihren Relevanzklassen benutzt. Dabei finden nur diejenigen relevanten Terme Berücksichtigung, deren Häufigkeit in der Datenbank ein gewisses Minimum überschreitet. Für die Tests wurde dieser Schwellwert auf drei gesetzt.
Diese Dokumentenbeschreibung berücksichtigt die Länge eines Dokuments, da diese in die Relevanzklasse jedes Terms eingeht. Allerdings enthalten längere Dokumente in

Folge der logarithmischen Skalierung der Relevanzklassen in der Regel mehr relevante Terme.
Typische Dokumente mit Längen zwischen 100 und 5000 Wörtern wie im Test verwendet - enthalten zwischen 10 und 100 relevante Terme.

Die Dokumentenbeschreibung ist formal einem sehr langen Dokumentvektor äquivalent, der aus natürlichen Zahlen besteht und dessen Komponenten den einzelnen Wörtern (bzw. ihren Grundformen) entsprechen, wobei die Komponentenwerte die Relevanzklassen der relevanten Terme sind. Ein solcher Vektor hat eine Länge der Größenordnung 10^6 (3×10^6 Terme in der Datenbank) und ist sehr schwach besetzt, da er nur etwa 10^1-10^3 von Null abweichende Komponenten aufweist.

5 Dokumentenähnlichkeit

Zur Definition der Dokumentenähnlichkeit wurden zwei Verfahren getestet, die nicht signifikant unterschiedliche Ergebnisse lieferten. Verglichen werden die zwei Dokumentvektoren $x=(x_i)$ und $y=(y_i)$:

5.1 Skalarprodukt

Hier wird das Skalarprodukt

$$sim_1(x,y)=S\ x_i\ y_i$$

der entsprechenden Dokumentvektoren verwendet, zwei Dokumente sind um so ähnlicher, je größer das entsprechende Skalarprodukt ist.

5.2 Durchschnitt der Menge der relevanten Terme

Statt über das Produkt wird das Minimum der Komponenten summiert:

$$sim_2(x,y)=S\ min(x_i\ ,y_i)$$

Die Dokumentenähnlichkeiten wird nur gespeichert, wenn sim_2 mindestens den Wert 4 liefert, da für kleinere Werte empirisch keine signifikante Ähnlichkeit festgestellt werden kann.

6 Dokumentenvergleich

Ein vollständiger paarweiser Dokumentenvergleich erfordert einen zur Dokumentenanzahl quadratischen Aufwand, was für große Kollektionen nicht praktikabel ist. Um den Aufwand gegenüber diesem naiven Ansatz zu verringern, wurde folgende Überlegung herangezogen: Die Dokumentenähnlichkeit kann nur dann von Null verschieden sein, wenn beide Dokumente wenigstens einen relevanten Term gemeinsam haben. Deshalb wird zunächst eine Liste aller relevanten Terme mit mehrfachem Vorkommen aufgestellt. Anschließend werden für jeden solchen Term ausschließlich die Dokumente paarweise auf Ähnlichkeit untersucht, die diesen Term als relevanten Deskriptor enthalten. Bezogen auf die Testkollektion von etwa 60.000 Dokumenten (s.o.) konnte so eine Laufzeitoptimierung um den Faktor 1000 erreicht werden.

Eine weitere Reduzierung ist möglich, wenn man die relevanten Terme auf ihre Produktivität hin bewertet und damit solche Terme aussondert, die zwar mehrfach als relevant auftreten, aber trotzdem nicht zur Ähnlichkeit beitragen.

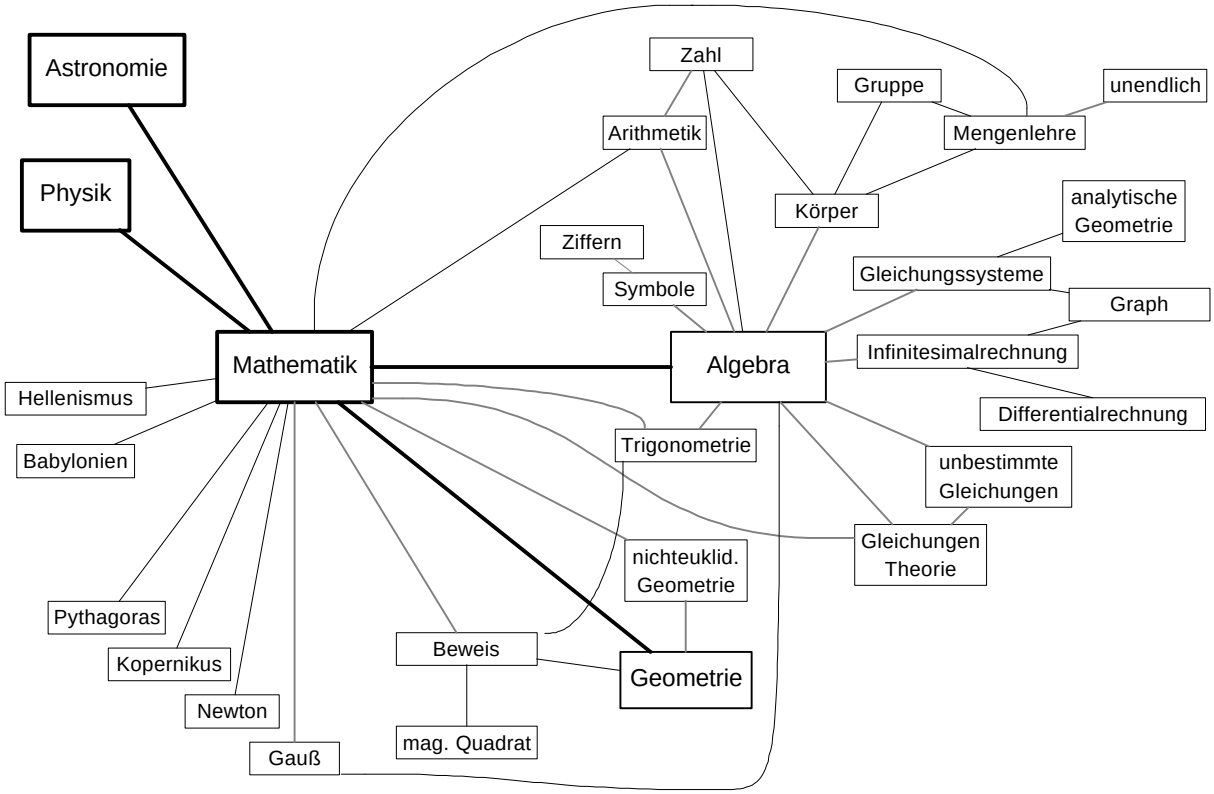
7 Ergebnisse

Im folgenden sollen zwei Beispiel-Cluster angegeben werden sowie statistische Aussagen über den Grad der Vernetzung getroffen werden. Dabei betrachten wir:

- Die Gesamtzahl der Verknüpfungen zwischen Dokumenten,
- die Verteilung der Verknüpfungsstärken und
- die Verteilung ausgehender Links.

7.1 Beispiel-Cluster

Im folgenden werden Dokumentcluster beschrieben, die ausgehend von einem Startdokument ermittelt wurden. Hervorgehobene Darstellung soll *Wichtigkeit der Dokumente* (d.h. viele ähnliche Dokumente) bzw. *Stärke der Ähnlichkeit* visualisieren . Die Anordnung erfolgte per Hand unter möglichst guter Berücksichtigung der Ähnlichkeiten. Eine ähnliche Anordnung sollte sich auch automatisch generieren lassen (vgl. Zizi 1996:210ff).



In Abb. 1 wird die Nachbarschaft von Begriffen im Umfeld des Stichwortes *Mathematik* aus Microsoft Encarta gezeigt. Eingezeichnet sind alle ermittelten Ähnlichkeiten der Stärke $sim_2 > 10$. Schwächere Ähnlichkeiten sind noch zahlreich vorhanden, sowohl zwischen den eingezeichneten Begriffen als auch zu weiteren Begriffen. Sämtliche gefundenen ähnlichen Artikel stammen ebenfalls aus Microsoft Encarta.

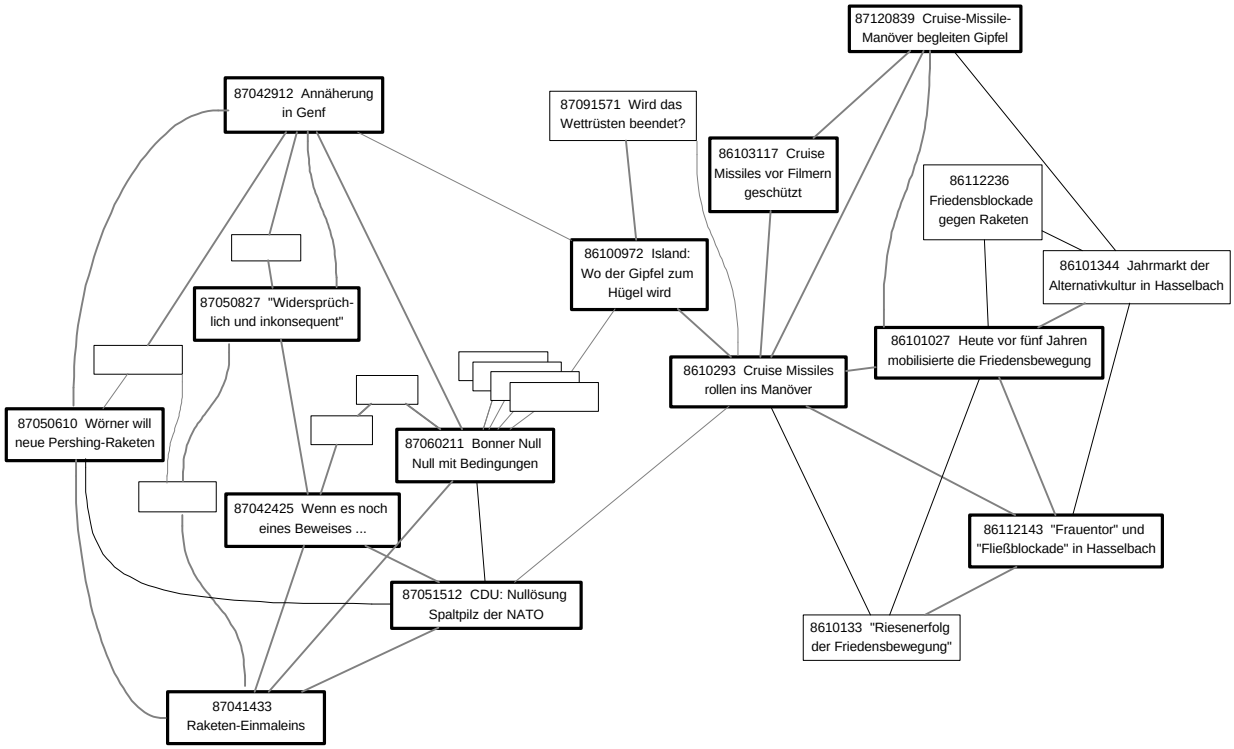


Abb. 2 zeigt zwei relativ schwach verbundene Cluster von Zeitungsartikeln zu dem Themenkomplex *Nachrüstung, Raketenstationierung, Friedensbewegung*. Dabei wird im rechten Teil der Zusammenhang von Artikeln zur Raketenstationierung in Deutschland sichtbar, wobei im Teil rechts unten der Schwerpunkt auf Aktivitäten der Friedensbewegung liegt, rechts oben dagegen bei Manövern mit Cruise Missiles in Deutschland. Im linken Teil liegt der Schwerpunkt bei der Diskussion um die NATO-Nachrüstung, wobei im Unteren Teil der Schwerpunkt Deutschland sichtbar ist. Der Übersichtlichkeit wegen wurden speziell im linken Teilcluster einige Dokumente und Links weggelassen.

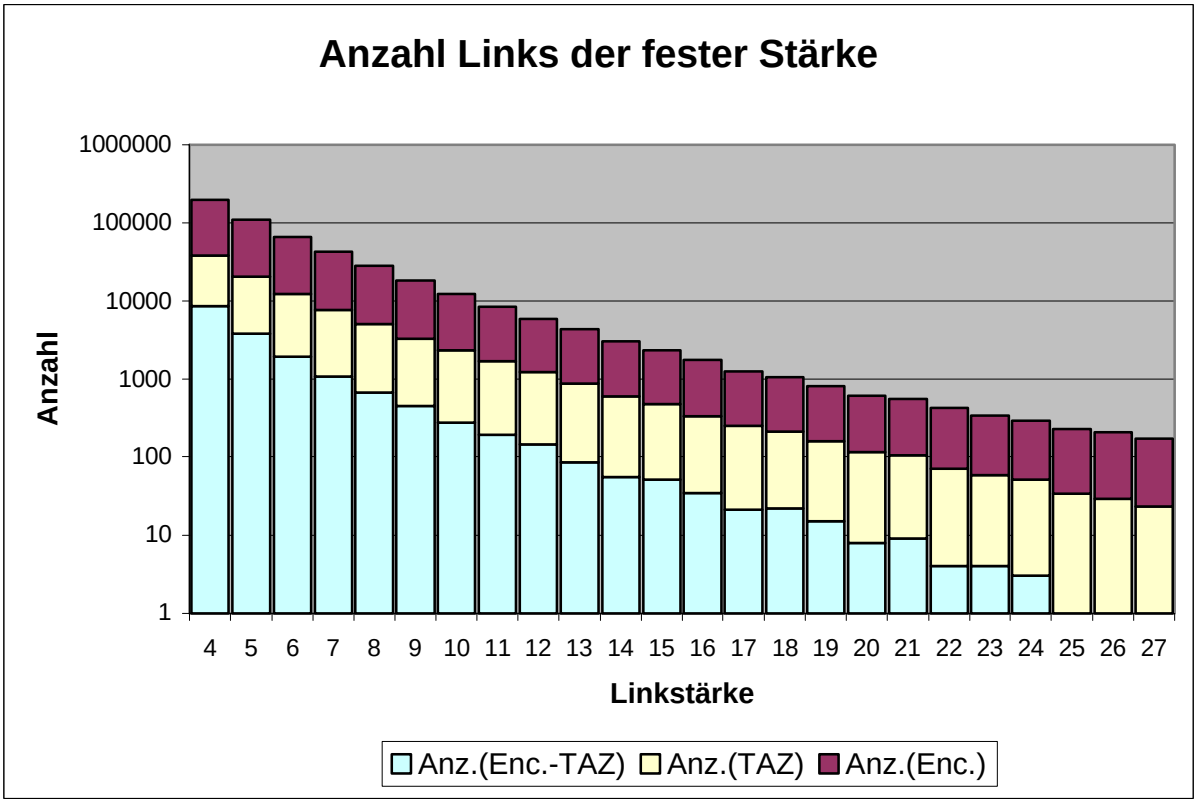
7.2 Gesamtzahl der Verknüpfungen

Insgesamt wurden 505011 Paare ähnlicher Dokumente ($sim_2 \geq 4$) gefunden. Dabei stammen bei 81,1% der Paare beide Dokumente aus Encarta, bei 15,4% der Paare beide Dokumente aus der TAZ und bei 3,5% der Paare jeweils ein Dokument aus Encarta und TAZ. Die durchschnittlichen Ähnlichkeiten sim_2 betragen 6,00, 6,05 und 5,36. Dies bestätigt die folgenden Unterschiede zwischen den Textsorten Lexikon und Zeitung:

- Beide Textsorten sind in sich konsistent (gleiche durchschnittliche Stärke der Verknüpfungen innerhalb der Klasse), aber voneinander verschieden. Deshalb gibt es zwischen den Textsorten extrem wenig Verknüpfungen, die zudem einen geringeren mittleren Ähnlichkeitswert haben.
- Zeitungsartikel haben häufig einen komplexeren Inhalt (z.B. Betrachtung des Zusammenhangs zweier Themen) als Lexikonartikel. Deshalb gibt es mehr Möglichkeiten für Inhalte und entsprechend weniger Verknüpfungen innerhalb des Zeitungskorpus.

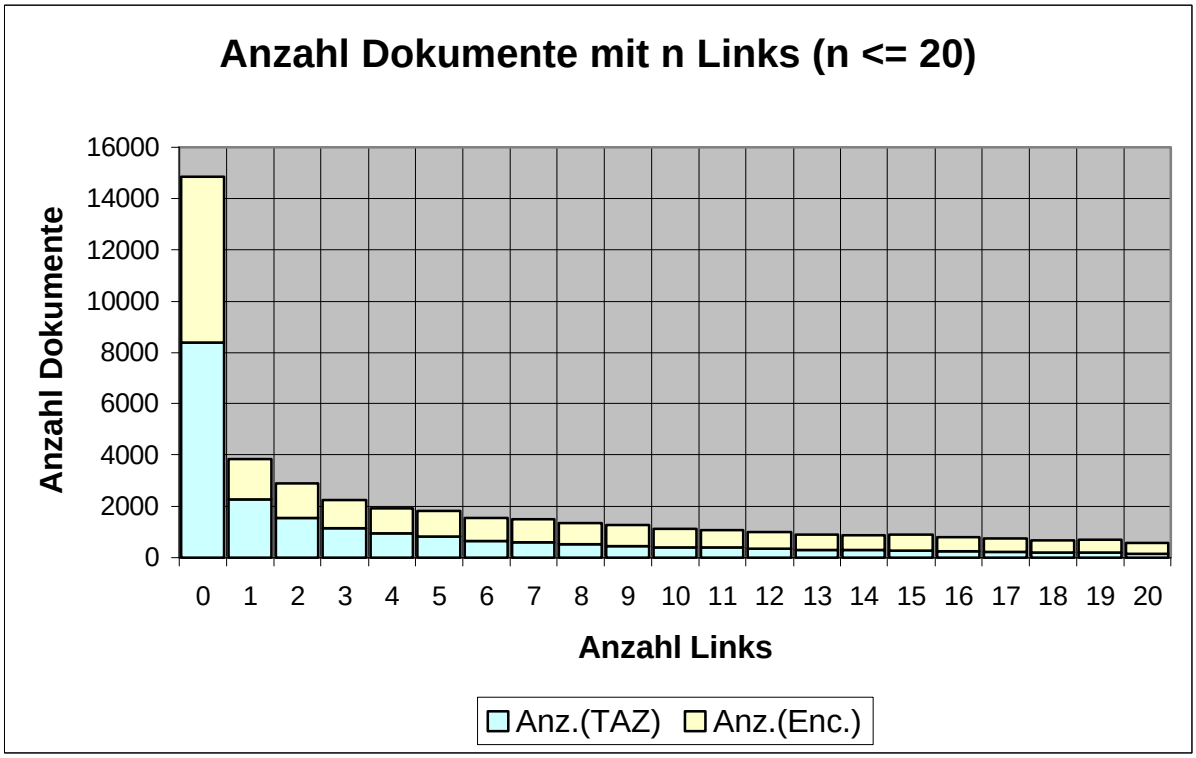
7.3 Verteilung der Verknüpfungsstärken

Abb. 3 zeigt, wieviel Dokumentenpaare mit Verknüpfungen einer bestimmten Stärke versehen sind versehen sind. Es zeigen sich keine typischen Differenzen zwischen den einzelnen Verteilungen. Berücksichtigt man die logarithmische Skalierung der Linkzahl, so zeigt die annähernd gleichmäßige Größe des oberen, dunklen Teils der Säulen einen konstanten Anteil der entsprechenden Links, d.h. eine annähernd konstante Verteilung der Links gleicher Stärke auf die Textsorten.



7.4 Verteilung der ausgehenden Links

Abb. 4 zeigt die Anzahl der Dokumente, von denen eine bestimmte Anzahl von Links ausgeht. Etwa ein Viertel der Dokumente sind nicht mit anderen verknüpft. Dies läßt sich teilweise mit sehr speziellen Themen erklären, zu denen es keine signifikanten Ähnlichkeitsbezüge gibt. Andererseits gibt es technische Schwierigkeiten, sehr kurze Dokumente zu verknüpfen, da diese über wenig signifikante Begriffe verfügen und das Verfahren das gemeinsame Auftreten von mindestens vier signifikanten Begriffen fordert.



8 Diskussion

Das beschriebene Verfahren simuliert das Verhalten des Menschen, die Ähnlichkeit von Objekten (hier: Dokumenten) an Hand ganz spezieller Eigenschaften (hier: niederfrequente Terme) zu erkennen. Im Gegensatz zum Menschen erfolgt hier allerdings keine zusätzliche Wertung dieser speziellen Eigenschaften darauf hin, inwieweit sie für die betrachtete Ähnlichkeit relevant sind, d.h. die Dokumentenbeziehungen sind nicht typisiert (vgl. AGOSTI et al. 1997:134f sowie ALLAN 1997), der versucht anhand formaler Kriterien eine derartige Linktypisierung für die automatische Hypertextgenerierung vorzunehmen). Eine Typisierung ist als Lernschritt im Rahmen eines Relevance-Feedback-Schrittes denkbar und sollte das Gesamtverhalten weiter verbessern.
Eine weitere Verbesserung ließe sich durch die Berücksichtigung von Mehrwortbegriffen erreichen. Dabei ist aber die vorbereitende Bereitstellung signifikanter Mehrwortbegriffe sowie der entsprechenden Häufigkeitsklassen in Analogie zu dem vorliegenden "Universalwortschatz" problematisch, da die dann auftretende kombinatorische Explosion mit derzeitigen Mitteln nicht beherrschbar erscheint.
Das geschilderte Verfahren kann in unterschiedlichen Anwendungsszenarien eingesetzt werden:

- Beim *information filtering*, wenn ausgehend von als relevant markierten Dokumente ähnliche Texte aus einem Dokumentenstrom selektiert werden sollen,
- als Unterstützung eines Recherche-Agenten (vgl. BRENNER et al. 1998:221ff), z.B. bei der Nachbearbeitung von Suchergebnissen oder einer lokalen Tiefensuche eines Web-Agenten oder
- für den automatischen Aufbau von Hypertexten aus einer Kollektion unverknüpfter Dokumente.

9 Literatur

AGOSTI, Maristella et al. (1997), On the Use of Information Retrieval Techniques for the Automatic Construction of Hypertext. In Information Processing & Management 33(2) (1997), 133-144.

ALLAN, James (1997). Building Hypertext Using Information Retrieval. In: Information Processing & Management 33(2) (1997), 145-159.

BRENNER, Walter et al. (1998). Intelligente Softwareagenten. Berlin et al.: Springer.

KORFHAGE, Robert R. (1996). Information Storage and Retrieval. New York et al.: John Wiley.

QUASTHOFF, Uwe (1998). Das Projekt Deutscher Wortschatz. In: HEYER, Gerhard; WOLFF, Christian (edd.) (1998). Linguistik und neue Medien. Proc. 10. Jahrestagung der Gesellschaft f. Linguistische Datenverarbeitung. Wiesbaden: Deustcher Universitätsverlag [erscheint].

SALTON, Gerard; MCGILL, Michael J. (1983). Introduction to Modern Information Retrieval. New York et al.: McGraw-Hill.

SALTON, Gerard (1989). Automatic Text Processing. The Transformation, Analysis, and Retrieval of Information by Computer. Reading/MA: Addison-Wesley.

SALTON, Gerard et al. (1990). "Automatic Structuring of Large Text Files." In: Communications of the ACM 37(2) (1994), 97-108.

SALTON, Gerard et al. (1996). Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts. In: AGOSTI & SMEATON (1996), 51-73.

ZIZI, Mountaz (1996). Interactive Dynamic Maps for Visuaklisation and Retrieval from Hypertext Systems. In: AGOSTI & SMEATON (1996), 203-224.