Softwaregestützte Analyse der Sichtbarkeit von Themen in internetbasierten Kommunikationsprozessen

Diplomarbeit im Fach Wirtschaftsinformatik

vorgelegt von

Peter Kiefer
Talblick 14a
91315 Höchstadt/A.
p.kiefer@f-t.de
Matrikelnummer: 1333053

10.03.2005

Angefertigt am
Lehrstuhl für Angewandte Informatik in den Kultur-, Geschichts- und Geowissenschaften

Fakultät Wirtschaftsinformatik und Angewandte Informatik
der
Otto-Friedrich-Universität Bamberg

Betreuung:
Prof. C. Schlieder

ZUSAMMENFASSUNG

Der Aufstieg des Internets zu einem weit verbreiteten und von großen Teilen der Bevölkerung intensiv genutzten Medium für jede Art der Kommunikation ermöglicht das Messen des Verlaufs von Themen im Bewusstsein der Öffentlichkeit mit Hilfe der Sichtbarkeit von Themen im Internet. Einfache Maßzahlen für eine solche Sichtbarkeit wie die Trefferzahl einer Suchmaschine berücksichtigen jedoch nicht die semantischen Beziehungen, die zwischen Themen bestehen können, und werden daher leicht missinterpretiert.

Ziel dieser Arbeit ist es, durch Modellierung der semantischen Beziehungen zwischen Themen und darauf aufbauender Anpassung der ursprünglichen Sichtbarkeitsmaße geeignetere Sichtbarkeitswerte zu berechnen.

Als Modellierungsansatz wird ein Thesaurus vorgeschlagen, aus dem nutzergestützt ein gerichteter und gewichteter Graph abgeleitet wird. Auf diesem Graph wird der Sichtbarkeitsausgleich algorithmisch durch einen um die Berücksichtigung der Co-Sichtbarkeit zwischen Themen ergänzten Spreading Activation-Algorithmus durchgeführt.

INHALTSVERZEICHNIS

1. EINLEITUNG

1.1 Sichtbarkeit von Themen im Fokus der Öffentlichkeit

Vergleicht man die Schlagzeilen verschiedener Tageszeitungen gleichen Datums oder Top-Meldungen diverser Nachrichtensendungen im Fernsehen, so fällt auf: Es sind im Allgemeinen die gleichen Themen, die sich zu einem bestimmten Zeitpunkt im Fokus der Medien befinden. Ereignisse (politischer, wirtschaftlicher, wissenschaftlicher, sportlicher oder auch kalendarischer Art) lösen ab einem gewissen – wie auch immer festgelegten – Interessantheitsgrad entsprechende Agenturmeldungen aus und führen zu einer Zunahme der Sichtbarkeit des zugeordneten Themas in den Medien.

So führen beispielsweise schlechte Ergebnisse bei einer PISA-Studie, der Konkurs eines großen Unternehmens, bahnbrechende Erkenntnisse in der Gentechnik, ein Dopingskandal im Sport oder einfach die Tatsache, dass Weihnachten vor der Tür steht, zu einer Häufung von Schlagzeilen oder Meldungen zu entsprechenden Themen.

Doch nicht nur eine Ähnlichkeit der aktuellen Themen in verschiedenen Aktoren (Zeitungen, Nachrichtensendungen) zu einem bestimmten Zeitpunkt lässt sich feststellen, auch ein zeitlicher Verlauf ist zu erwarten: Mit abnehmendem Neuheitsgrad des Themas wird das Interesse der Öffentlichkeit sinken und neue Ereignisse werden andere Themen in den Mittelpunkt rücken, so dass die Sichtbarkeit eines Themas im Laufe von Tagen oder Wochen natürlicherweise abnehmen wird. Weiterhin ist denkbar, dass verschiedene Themen in einem Kausalzusammenhang stehen und daher mit hoher Wahrscheinlichkeit zeitgleich diskutiert werden (z.B. „PISA-Studie“ und „Bildungspolitik“) oder zeitlich aufeinander folgen (z.B. „Bildungspolitik“ → „Föderalismusreform“).

Nicht unerwähnt bleiben soll an dieser Stelle die Möglichkeit, dass die Medien nicht nur berichtende Funktion übernehmen, sondern ihrerseits durch Fokussierung auf bestimmte Themen Ereignisse (z.B. politischer Art) auslösen und somit auch gestaltende Funktion übernehmen können (vgl. z.B. [NeiEilPfe98]).

Es ist durchaus möglich, den oben beschriebenen Verlauf der Sichtbarkeit von Themen in Printmedien und Fernsehen zu analysieren. Die dafür notwendige täglich durchzuführende manuelle Inhaltsextraktion aus einer umfangreichen Menge in Papierform vorliegender Texte (bzw. die Analyse von Videomaterial) ist sehr aufwendig und bleibt daher im Umfang meist auf einige wenige Medien beschränkt.

1.2 Sichtbarkeit von Themen in internetbasierten Kommunikationsprozessen

Die Idee dieser Arbeit ist es, die in 1.1 beschriebene Analyse der Sichtbarkeit von Themen nicht in herkömmlichen Printmedien oder dem Fernsehen, sondern in internetbasierten Kommunikationsprozessen durchzuführen. Das Internet eignet sich hierfür aus mehreren Gründen:

a) Die zunehmende Bedeutung des Internets auf der Schwelle vom Industrie- zum Informationszeitalter[1].
b) Die ständige Aktualisierung der Internetseiten ermöglicht eine sehr zeitnahe Analyse der Veränderungen von Sichtbarkeiten auf Grund bestimmter Ereignisse.
c) Relativ geringer Aufwand beim Ermitteln der Sichtbarkeitsdaten durch Automatisierung und Vorliegen der Dokumente in elektronischer Form.
d) Das Überwachen mehrerer Aktoren (z.B. Domains) kostet nicht wesentlich mehr Zeit und Aufwand als das Überwachen eines Aktoren (i.Ggs. zum Überwachen mehrerer Zeitungen). Entsprechendes gilt für das Monitoring über einen längeren Zeitraum.

[1] Das rasante Wachstum des Internets, die allmählich zu erwartende Verdrängung herkömmlicher News-Medien durch e-paper und Newsticker, die essentielle Bedeutung der Onlinepräsenz für die Unternehmenskommunikation, die zunehmende Rolle von Internetforen und Webtagebüchern zur gesellschaftlichen Meinungsbildung sowie die vielen anderen Phänomene der „Erfolgsgeschichte Internet“ sollen hier nicht weiter diskutiert werden.

e) Eine globale Suche ist möglich. Ein Beobachten aller Tageszeitungen oder aller Nachrichtensendungen hingegen ist kaum machbar.

f) Der Zugriff auf Dokumente, die nur in elektronischer Form vorliegen oder die ohne Internet nur schwer zugänglich wären, wird ermöglicht.

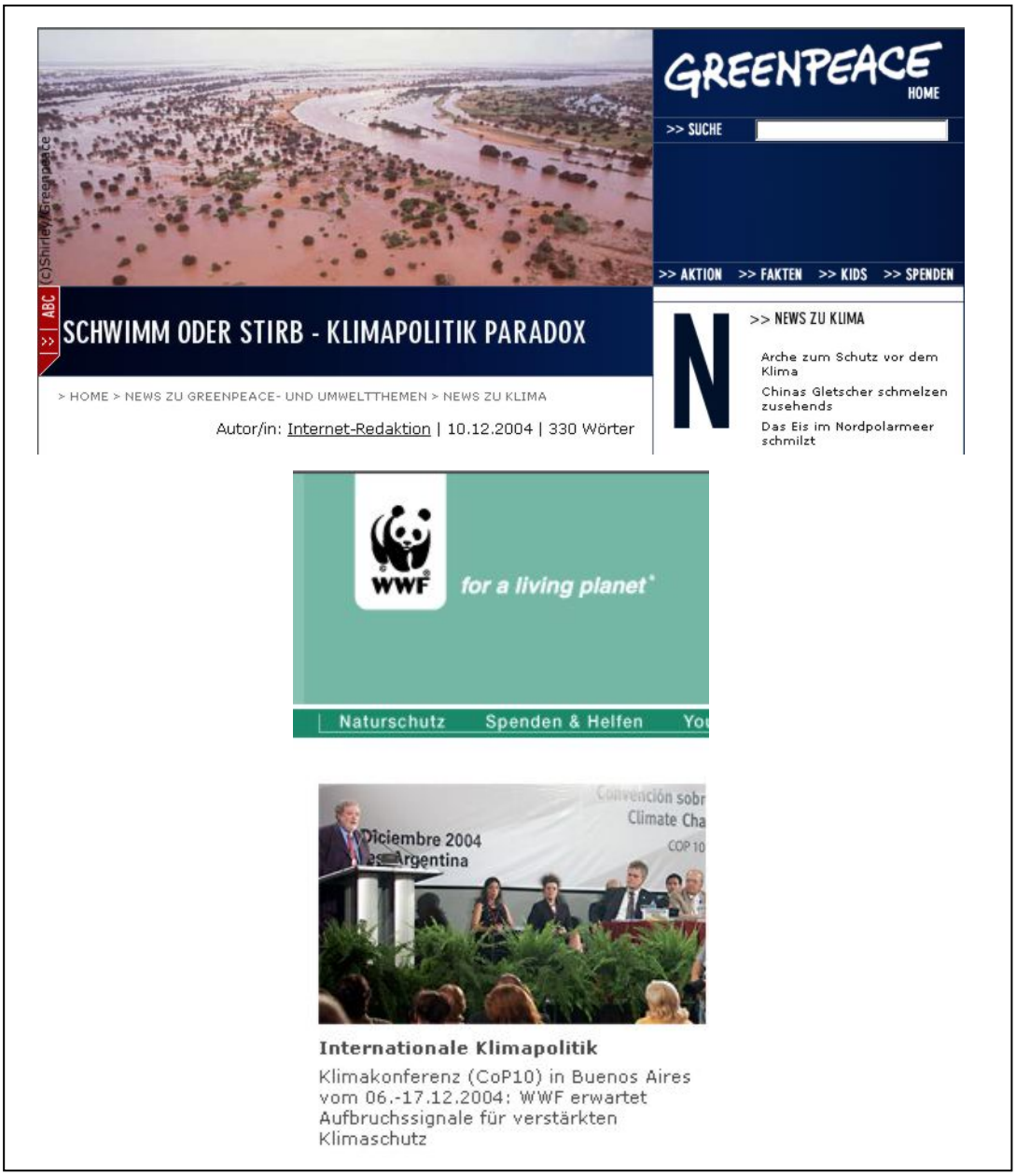


Abbildung 1: Startseiten von www.greenpeace.de und www.wwf.de am 10.12.04

Abbildung 1 zeigt die Startseiten von www.greenpeace.de und www.wwf.de am 10.12.2004. Beide thematisieren „Klimapolitik" – auf beiden Seiten ist die Sichtbarkeit dieses Themas zum gleichen Zeitpunkt sehr hoch. Es kann vermutet werden, dass die Klimakonferenz in

Buenos Aires Anlass hierfür war, was ein Indiz dafür ist, dass Ereignisse nicht nur Spuren in der Berichterstattung in TV und Zeitung hinterlassen, sondern auch eine Auswirkung im Internet haben.

Beobachtete man – wie hier auf www.greenpeace.de und www.wwf.de geschehen – die Themen im Internet per Hand, so wäre der ersparte Aufwand sicherlich gering. Daher sollte die Sichtbarkeitsanalyse von Themen im Internet unter Nutzung der oben genannten Vorteile von einem Tool unterstützt werden. Im Rahmen dieser Arbeit entsteht ein solches Tool in Form eines Java-Programms.

Doch wer hätte an einer solchen toolgestützten Analyse der Sichtbarkeit von Themen im Internet Interesse? Im Folgenden sollen drei mögliche Interessengruppen und zugleich beispielhafte Anwendungsfälle vorgestellt werden.

1.3 Mögliche Interessengruppen

1.3.1 Unternehmenskommunikation

Der Wandel von der Industrie- zur Informationsgesellschaft geht einher mit zwei Tendenzen: Einerseits einer kaum zu bewältigenden Informationsflut, die das Individuum vor Selektions- und Bewertungsprobleme bei der Suche nach relevanter Information stellt, während die hierzu von ihm benötigte Aufmerksamkeit durch Zeitrestriktion und kognitive Grenzen beschränkt ist. Andererseits wird es durch diese allgemeine Informationsüberflutung für Unternehmen (aber auch Non-Profit-Organisationen) zunehmend schwieriger, die Aufmerksamkeit ihrer Zielgruppe auf sich zu lenken. So deklariert Franck, Aufmerksamkeit sei „die knappste Ressource der Informationsverarbeitung“ und erhebt sie sogar in den Rang einer „neue[n] Währung“ ([Franck99], S.8; vgl. hierzu auch weitere Literatur zur Aufmerksamkeitsökonomie, z.B. [Falkinger03], [Goldhaber97]).

Die Aufmerksamkeit der Öffentlichkeit für ein Thema, das von einem Unternehmen (einer Organisation/einer Partei) in das allgemeine Bewusstsein gerückt werden soll, kann unter Umständen an Hand der Sichtbarkeit dieses Themas im Internet gemessen werden: In Foren

wird verstärkt über das Thema diskutiert, Online-News enthalten dieses Thema, Produkte des Unternehmens werden vermehrt in Online-Shops angeboten. Somit kann eventuell der Erfolg einer Kampagne mit Hilfe des zeitlichen Verlaufs der Sichtbarkeit des Themas beurteilt werden. Freilich ist ein Kausalzusammenhang zwischen der Öffentlichkeitsarbeit und zunehmender Sichtbarkeit im Internet nicht zwangsweise gegeben, da das Thema auch aus anderen Gründen die öffentliche Aufmerksamkeit auf sich gezogen haben könnte. Außerdem könnte eine Stagnation der Sichtbarkeit lediglich bedeuten, dass die Sichtbarkeit ohne die öffentlichkeitswirksamen Maßnahmen gesunken wäre.
Ein weiterer Nutzen besteht in der Untersuchung der Nachhaltigkeit von Öffentlichkeitsarbeit an Hand der Geschwindigkeit der Abnahme der Sichtbarkeit eines Themas nach Ablauf der Kampagne. Aus Sicht eines multinationalen Unternehmens ergibt sich weiterhin die Möglichkeit der Analyse der Sichtbarkeit eines Themas in verschiedenen Testmärkten durch Differenzierung der Sichtbarkeitsanalyse nach Domains (z.B. „de“/„at“).

Beispielsweise könnte eine Bürgerrechtsorganisation beschließen, im Monat März das Thema „genetischer Fingerabdruck“ zu forcieren, und die Sichtbarkeit dieses Themas von Februar bis Mai messen. Sollte die Sichtbarkeit nun Mitte März zunehmen, kann das im besten Fall an der Öffentlichkeitsarbeit dieser Organisation liegen, evtl. aber auch an einem spektakulären Fahndungserfolg der Polizei mit Hilfe des genetischen Fingerabdrucks, der für Schlagzeilen sorgt.

Ein aktuelles Beispiel aus dem Marketing ist die Werbekampagne der Alcina® Cosmectic Dr. Kurt Wolff® GmbH & Co KG Anfang 2005 (Markenname Alpecin®) für ihren neuen „Coffein-Complex“, der im „After Shampoo Liquid“® dem Haarausfall vorbeugen soll. In den Werbespots wird für den „Glatzenrechner“ auf der Homepage des Unternehmens geworben, der einem nach nur wenigen Benutzereingaben das wahrscheinliche Alter der vollständigen Glatzenbildung errechnet.
Von den ersten TV-Werbespots im Januar bis 06.02.2005 beobachtet ergibt sich der in Abbildung 2 dargestellte Verlauf der geschätzten Trefferzahlen für „Coffein-Complex“ (linke Achse) und „Glatzenrechner“ (rechte Achse) von Google™ (zum Messen der Sichtbarkeit auf Webseiten vgl. 2.2.1): Die Werte von „Coffein-Complex“ steigen von knapp über Null bis auf 48 an, der „Glatzenrechner“ steigt von Werten um die vierhundert bis über die Tausender-

marke. Die Beurteilung von Erfolg oder Misserfolg der Werbekampagne bleibt den Marketingspezialisten überlassen.

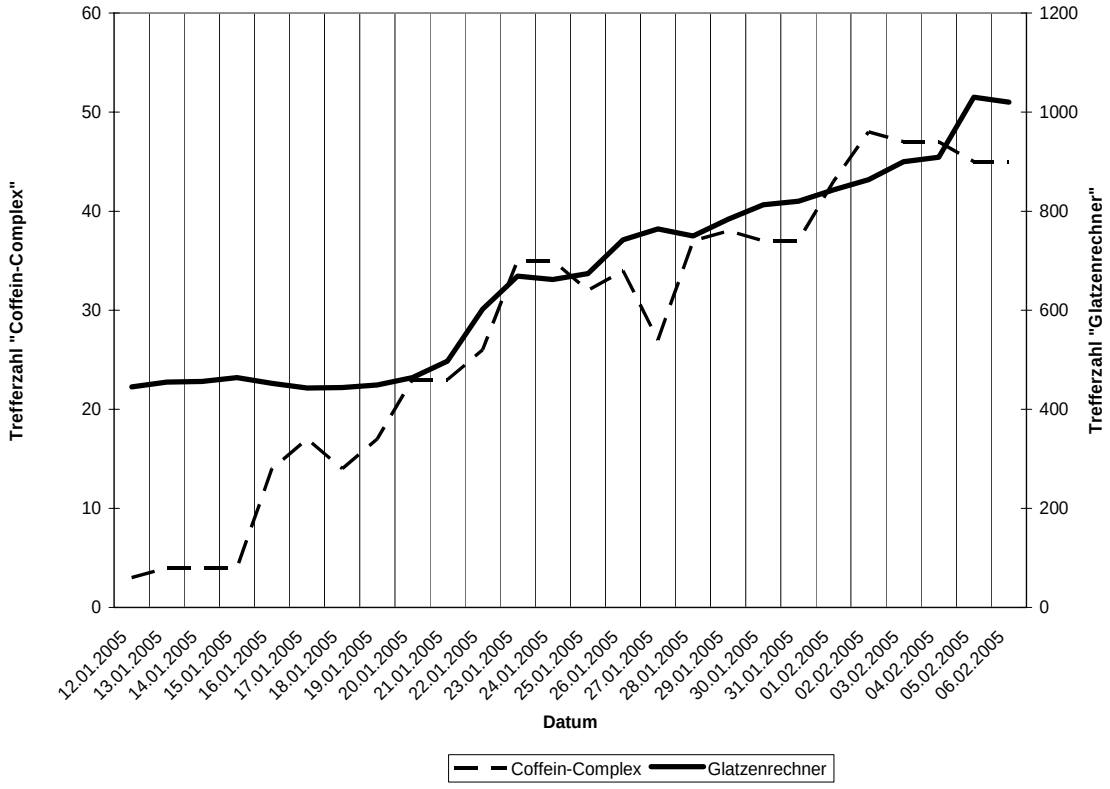


Abbildung 2: Verlauf der Trefferzahlen der Suchmaschine Google™ für „Coffein-Complex“ und „Glatzenrechner“ vom 12.01.05 bis 06.02.05

1.3.2 Administratoren von Internetforen

Im Gegensatz zu der unidirektionalen 1:n-Kommunikation der Unternehmenskommunikation ermöglichen Diskussionsforen eine Kommunikation zwischen beliebig vielen Teilnehmern, wobei jeder Teilnehmer gleichzeitig die Rolle eines Senders und eines Empfängers von Nachrichten einnimmt. Internetforen existieren in diversen Formen: Moderiert oder unmoderiert, offen oder mit eingeschränktem Teilnehmerkreis, welche mit der Linkstruktur eines Baumes oder eines Graphen, sowie Foren zu den verschiedensten Oberthemen.

Zu einem solchen Oberthema – z.B. „Fußball Weltmeisterschaft in Deutschland 2006“ – lassen sich meist mehrere Unterthemen identifizieren – z.B. „Ticketvorverkauf“, „Hooligans“, „Reisebeschränkungen“ – deren relative Sichtbarkeit zueinander sich im Zeitverlauf verändert. Im Beispiel ist sicherlich „Ticketvorverkauf“ im Jahr 2005 sehr sichtbar, während „Hoo-

ligans“ eher während der WM im Jahr 2006 diskutiert wird. Als Themenabfolge könnte festgestellt werden, dass das Thema „Reisebeschränkungen“ aus dem Thema „Hooligans“ hervorgeht.

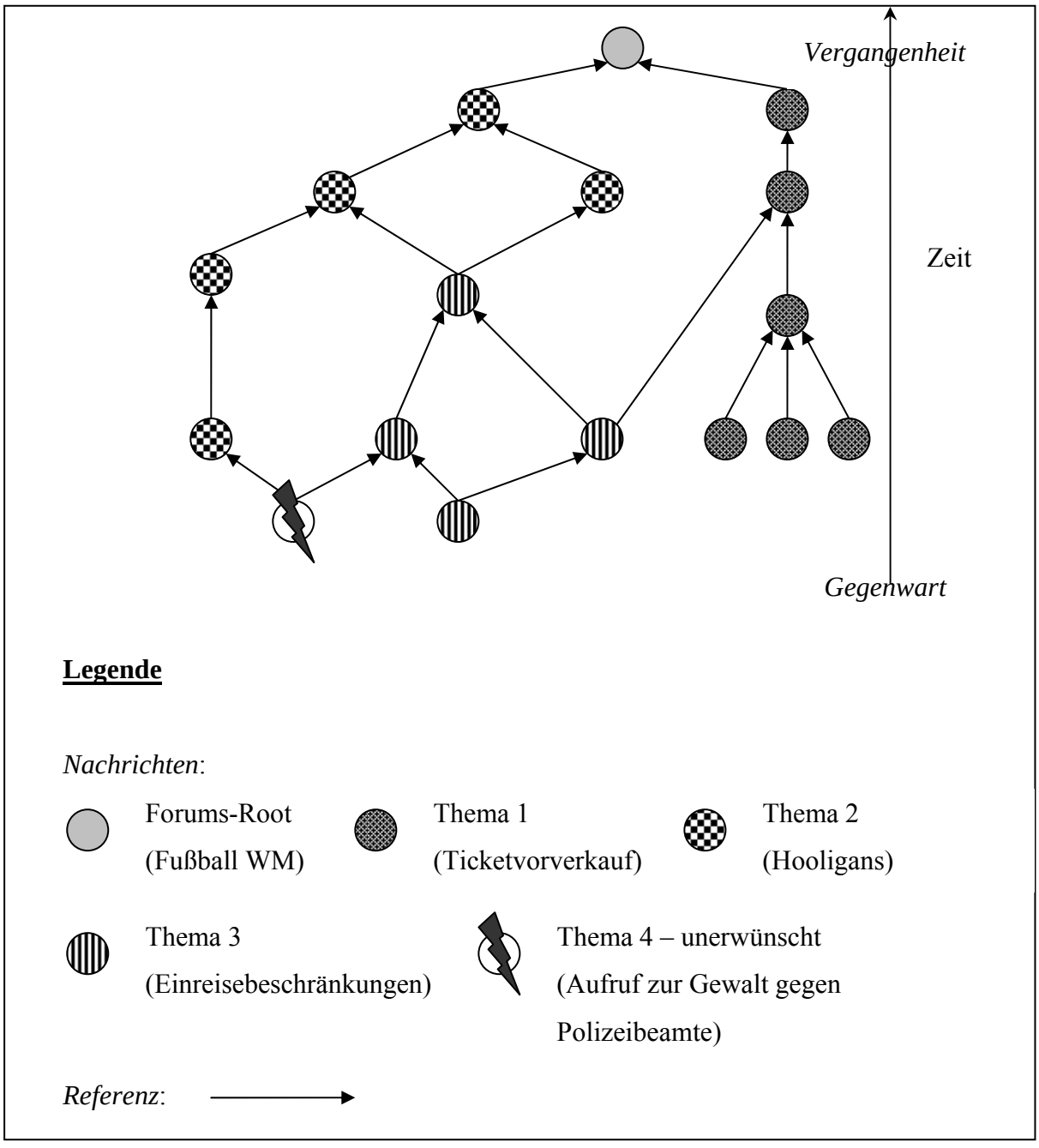


Abbildung 3: Referenzstruktur zwischen Nachrichten mit zugeordneten Themen in einem Internetforum

Ein Moderator eines solchen Forums hat sicherlich Interesse am Verlauf der Diskussion über die diversen Unterthemen, ohne täglich sämtliche neuen Nachrichten lesen zu wollen, und

wird im Bedarfsfall eingreifen müssen, um die Diskussion in eine andere Richtung zu lenken, falls beispielsweise die Diskussion zum Thema „Hooligans“ in Aufrufe zu Straftaten ausartet. Zusätzlich wäre eine geeignete Visualisierung der Zuordnung von Themen zu Nachrichten wünschenswert. Abbildung 3 stellt das beschriebene Beispiel dar: Ein Forum mit Nachrichten, die zu verschiedenen Zeitpunkten entstehen und auf Nachrichten vergangener Zeitpunkte verweisen (gerichteter azyklischer Graph). Die Nachrichten werden auf Grund des in ihnen hauptsächlich diskutierten Themas gekennzeichnet. Die Themenübergänge lassen sich dadurch ebenso beschreiben wie unerwünschte Nachrichten aufspüren.
Zum Messen der Sichtbarkeit in Internetforen vgl. 2.2.2.

1.3.3 Sozialwissenschaftler

Auch das Messen gesellschaftlicher Strömungen, die beispielsweise für Soziologen oder Politologen von Interesse sind, kann durch eine Sichtbarkeitsanalyse im Internet unterstützt werden. Solche Untersuchungen sind meist mittel- oder langfristiger Art und sollten durch eine eingehende Analyse der Ergebnisse begleitet werden. Von besonderem Interesse dürfte hier nicht sein, dass, sondern aus welchem Grund sich die Sichtbarkeit eines Themas verändert hat, wobei für eine solche Erklärung ein Hinzuziehen von Informationen aus anderen Quellen sowie ein tiefes Verständnis der Fachrichtung unerlässlich ist. Diese Erklärungsversuche bleiben den Spezialisten des jeweiligen Faches überlassen.

Ein mögliches Beispiel wäre der Vergleich der Sichtbarkeitsverläufe des Themas Aids im gesamten Internet, in Internetforen für Jugendliche und in solchen für homosexuelle Männer. Aufgabe des Forschers ist es, anschließend daraus Rückschlüsse auf das Bewusstsein für die Aids-Problematik in verschiedenen gesellschaftlichen Gruppen zu ziehen. Eine plötzliche Zunahme oder Abnahme der Sichtbarkeit kann außerdem bei entsprechender Fachkenntnis mit jeweils aktuellen Ereignissen in Verbindung gebracht werden (z.B. 1. Dezember: Weltaidstag).

Im folgenden Diagramm (Abbildung 4) wird nicht der Verlauf von Sichtbarkeiten, sondern das Verhältnis von Trefferzahlen verschiedener im Zusammenhang mit HIV und Geschlechtskrankheiten stehender Themen zueinander in verschiedenen Domains zu einem bestimmten Zeitpunkt (01.02.2005) dargestellt.

Neben der Datenreihe für keine Domaineinschränkung (*) wurden als Domains ein Forum für Jugendliche (www.oxybrain.de), eine Seite für Homosexuelle (www.eurogay.de) sowie zwei Seiten von aufklärenden Organisationen (Bundeszentrale für gesundheitliche Aufklärung, www.bzga.de; www.aidshilfe.de) gewählt. Zu beachten ist an dieser Stelle, dass es sich hier um Prozente handelt, die nichts über die absolute Anzahl der Webseiten in den einzelnen Domains aussagen. Außerdem werden lediglich die untersuchten Themen miteinander verglichen – im Jugendforum z.B. werden sicherlich noch viele andere Themen diskutiert, die hier gar nicht berücksichtigt werden. Aus der hier dargestellten Graphik können somit lediglich Aussagen der Art „Auf www.eurogay.de ist das Thema ‚Analverkehr' sichtbarer als das Thema ‚Kondom'" getroffen werden.

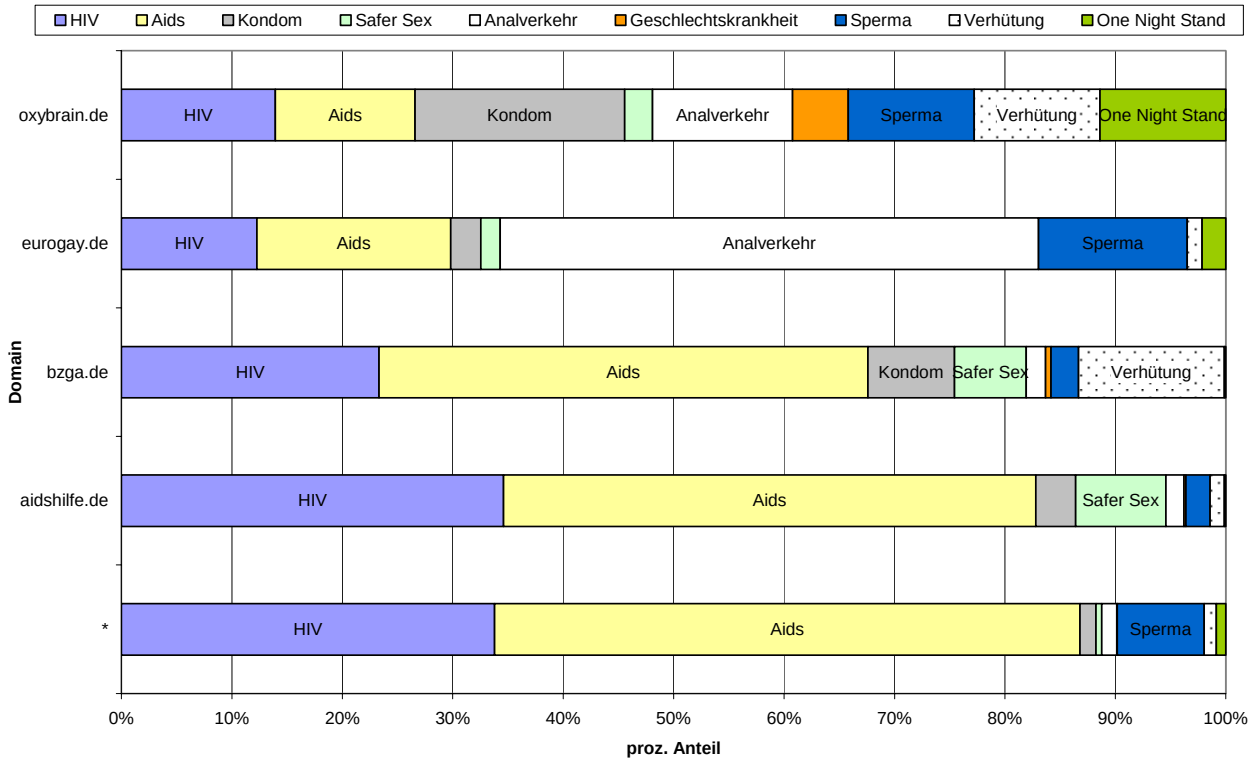


Abbildung 4: Verhältnisse der Trefferzahlen (Suchmaschine Google™) verschiedener Themen aus dem Themengebiet HIV zueinander (fünf verschiedene Domains)

1.4 Zielsetzung und Aufbau dieser Arbeit

Der folgende Abschnitt 2 gibt eine Einführung in die Werkzeuge der Sichtbarkeitsanalyse und zeigt, wie Sichtbarkeitsdaten für Themen aus zwei beispielhaften internetbasierten Kommunikationsprozessen gewonnen werden können. Auf die Verwendung eines durchgehenden Praxisbeispiels wird hierbei (wie auch in Abschnitt 1) bewusst verzichtet, um die vielseitige Einsetzbarkeit der Sichtbarkeitsanalyse zu illustrieren. Auch die Idee der gemeinsamen Sichtbarkeit (Co-Sichtbarkeit) zweier Themen wird am Ende von Abschnitt 2 eingeführt.

Abschnitt 3 legt dar, dass die mit den einfachen Methoden aus Abschnitt 2 gewonnenen Sichtbarkeitsdaten in vielen Fällen zu gravierenden Missinterpretationen führen. Ziel dieser Arbeit ist es, einen Ansatz zur nachträglichen Anpassung der Sichtbarkeitsdaten zu finden und somit zutreffendere Sichtbarkeitsdaten für Themen zu erhalten.

Dieser Ansatz wird in Abschnitten 4 und 5 erarbeitet. Hierbei beschäftigt sich Abschnitt 4 mit einem Teilproblem der semantischen Modellierung der Beziehungen zwischen Themen, während Abschnitt 5 darauf aufbauend einen Algorithmus zur Veränderung der Sichtbarkeitsdaten vorschlägt.

Erwähnenswerte Aspekte, die bei der im Rahmen dieser Arbeit entstandenen Realisierung eines Tools zur Sichtbarkeitsanalyse auffielen, werden in Abschnitt 6 kurz aufgeführt. Abschnitt 7 evaluiert schließlich den gewählten Lösungsansatz und mündet in Abschnitt 8 – eine Zusammenfassung mit Ausblick auf noch zu lösende Aspekte der Sichtbarkeitsanalyse.

2. ANALYSE DER SICHTBARKEIT VON THEMEN

Bevor ab Abschnitt 3 auf die spezielle Problemstellung dieser Arbeit eingegangen wird, sollen in 2 einige zum Verständnis notwendige Grundideen eingeführt werden:

- Wie lässt sich der bisher intuitiv verwendete Begriff der Sichtbarkeit eines Themas allgemein definieren?
- Was sind internetbasierte Kommunikationsprozesse, die für eine solche Analyse in Frage kommen, und wie lässt sich ein Sichtbarkeitsmaß für diese individuellen Fälle definieren?
- Wie lässt sich die gemeinsame Sichtbarkeit zweier Themen beschreiben?

2.1 Allgemeine Definition eines Sichtbarkeitsmaßes

Im Folgenden sollen einige Begriffe definiert und Bezeichnungen eingeführt werden, die für den Rest dieser Arbeit von Belang sind.

TOP

bezeichnet die Menge aller Themen.

In dieser Arbeit wird ein Thema lediglich durch seinen Namen identifiziert und hinreichend beschrieben (z.B. „Klimaschutz“); $top \in TOP$ ist also im Folgenden der Einfachheit halber stets eine Zeichenkette. Dabei sollte nicht übersehen werden, dass in konkreten Anwendungsfällen die genaue Abgrenzung eines Themas notwendig ist und eine eventuelle zusätzliche Zuordnung von Eigenschaften möglich sein sollte. Dies bleibt den Anwendern der hier beschriebenen Methoden als Spezialisten auf ihrem jeweiligen Fachgebiet überlassen.

DOCUMENTS

bezeichnet die Menge aller untersuchten Dokumente.

relevance: $TOP \times DOCUMENTS \rightarrow [0..1]$

ist eine Relevanzfunktion, die den Dokumenten einen Relevanzwert für jedes Thema zuordnet. Hierzu sind im Allgemeinen Methoden des Information Retrieval anzuwenden (vgl. z.B. [BaezaRib04]).

structure

ist Strukturinformation über DOCUMENTS. Diese kann beispielsweise eine Linkstruktur im Sinne eines allgemeinen Graphen beschreiben. Auch ein Fehlen der Strukturinformation ist möglich, wenn DOCUMENTS als Menge ungeordneter Dokumente aufgefasst wird.

visibility$_{\textbf{relevance, structure, DOCUMENTS}}$**:** $TOP \rightarrow IR^{+}_{0}$

schließlich bezeichnet die Sichtbarkeitsfunktion, die jedem Thema $top \in TOP$ einen positiven Sichtbarkeitswert zuordnet - abhängig von den vorhandenen Dokumenten DOCUMENTS, der Relevanzbewertung relevance und der Strukturinformation structure. Ein hoher visibility-Wert steht hierbei für eine hohe Sichtbarkeit, 0 für gar keine Sichtbarkeit.

2.2 Beispiele typischer internetbasierter Kommunikationsprozesse

2.2.1 Webseiten

Der einfachste und unpersönlichste Kommunikationsprozess im Internet besteht in der simplen Veröffentlichung von Webseiten. Unpersönlich, da sich Internetseiten grundsätzlich nicht an eine bestimmte Person oder meist auch nicht an einen genau begrenzbaren Personenkreis richten, sondern an unbekannte Adressaten, die aus einem beliebigen Grund Interesse an der Seite haben.

2.2.1.1 Sichtbarkeitsmaß für Themen auf Webseiten

Ein sehr einfaches Sichtbarkeitsmaß für Themen auf Webseiten ergibt sich aus der in 2.1 beschriebenen Definition wie folgt:

DOCUMENTS ist in diesem Fall eine Menge von Web-Dokumenten (meist HTML).

Denkbar ist die Verwendung aller im Web vorhandener Dokumente oder auch die Selektion der Dokumente anhand ihrer Domain (z.B. alle Dokumente der Domain greenpeace.org). Dies ermöglicht den Vergleich von Sichtbarkeiten in verschiedenen Aktoren (z.B. Vergleich der Sichtbarkeit eines Themas auf www.greenpeace.org und www.wwf.de).

relevance werde für den einfachsten Fall wie folgt definiert:

relevance(top, doc) = 0, falls das Dokument doc das Thema top nicht enthält[1]
relevance(top, doc) = 1, falls das Dokument doc das Thema top enthält

Anmerkung: Eine komplexere relevance-Funktion könnte die Häufigkeit des Auftretens von top oder beispielsweise die Stelle des Auftretens berücksichtigen (z.B. höhere Bewertung von top in Überschriften/Header-Tags).

structure ist in diesem einfachsten Fall leer, d.h. keine Strukturinformation vorhanden. DOCUMENTS ist eine ungeordnete Menge von HTML-Dokumenten.

Anmerkung: Eine komplexere Definition würde die Linkstruktur zwischen den Dokumenten berücksichtigen. Die visibility-Funktion könnte anschließend diese Strukturinformation nutzen, um Seiten, die leichter über Links zu erreichen sind, eine höhere Gewichtung zu verschaffen.

visibility(top) wird schließlich in diesem einfachsten Fall als die Anzahl aller Dokumente definiert, für die relevance(doc) = 1 gilt.

$$\text{visibility}(top) = \sum_{doc \in DOCUMENTS} \text{relevance}(top, doc)$$

[1] wie in 2.1 erwähnt steht top hier für eine Zeichenkette

D.h. die Sichtbarkeit eines Themas ergibt sich einfach aus der Anzahl der Webseiten, die dieses Thema enthalten.

Diese Sichtbarkeitsdefinition ist zwar sehr grob, kann jedoch technisch relativ einfach über den Recall-Wert einer Suchmaschine realisiert werden. Als Sichtbarkeit für ein Thema wird die (evtl. geschätzte) Trefferzahl einer Suchmaschine für dieses Thema verwendet. Da im Rahmen dieser Arbeit diese Daten nur als initiale Sichtbarkeiten und somit als Inputdaten betrachtet werden, soll diese einfache Definition an dieser Stelle genügen.

2.2.2 Diskussionsforen

Eine weitere nicht mehr aus dem Internet wegzudenkende Kommunikationsform sind die bereits in 1.3.2 erwähnten Diskussionsforen, die hier als zweiter internetbasierter Kommunikationsprozess erwähnt werden sollen. Das im Folgenden in 2.2.2.2 definierte Sichtbarkeitsmaß für Themen in Diskussionsforen basiert auf einem Sichtbarkeitsmaß für Nachrichten. Ein solches wurde im Rahmen des COM-Ansatzes (Communication-Oriented Modeling) formalisiert und wird zunächst kurz vorgestellt.

2.2.2.1 Sichtbarkeitsmaß für Nachrichten in Diskussionsforen (COM)

Communication-Oriented Modeling (COM, vgl. zu diesem Punkt [MaSchlie02]) betrachtet internetbasierte Kommunikationsprozesse mit vielen Teilnehmern nicht auf Grund von Agent-Agent-Beziehungen (Agent-Oriented Modeling), sondern die Beziehungen zwischen den Nachrichten. Diese als persistent betrachteten Nachrichten werden somit nicht an einen bestimmten Agenten gerichtet, sondern an die Allgemeinheit – „To Whom It May Concern“ (vgl. [MaSchlie02], S. 1) – und enthalten Referenzen auf ältere Nachrichten, auf die sie sich thematisch beziehen.
Formal gesehen handelt es sich um eine Halbordnung über „publication events“, mit der die zeitliche Ordnung der Nachrichten festgelegt wird (jeder Nachricht ist bijektiv genau ein „publication event“ zugeordnet), und eine zusätzliche binäre Relation auf den Nachrichten, die die Referenzstruktur der Nachrichten zum Ausdruck bringt. Letztere muss hierbei passend zu der zeitlichen Ordnung sein, d.h. keine Nachricht darf auf eine Nachricht gleicher oder aktuellerer Zeit verweisen (vgl. Abbildung 5).

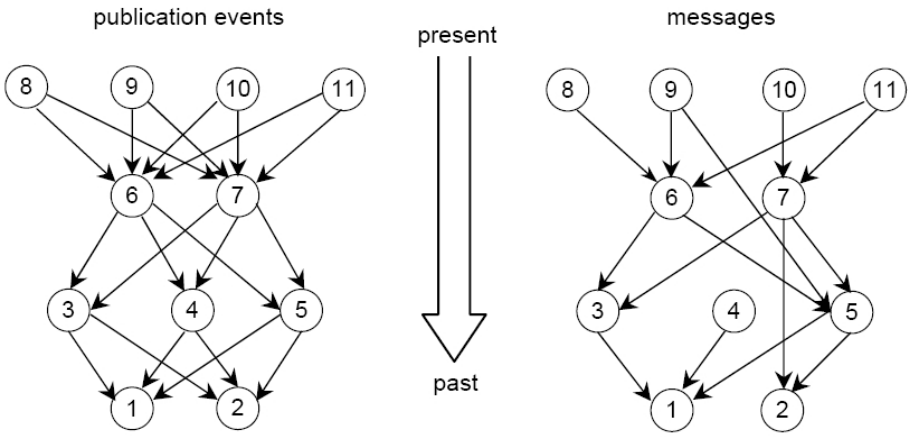


Abbildung 5: zeitliche Halbordnung (Hasse Diagramm) und Referenzen zwischen Nachrichten (vgl. [MaSchlie02], Fig. 3)

Die Sichtbarkeit einer Nachricht spielt im COM-Ansatz eine wesentliche Rolle:

> *„Even with technically persistant messages, a decrease in accessibility will occur over time because the access to a message is linked to its social visibility in the communication process. The tendency of messages to become less visible over time is counterbalanced by the tendency of references to increase the visibility of the message that is referred.“* (vgl. [MaSchlie02], S. 12)

Die Sichtbarkeit einer Nachricht, die als eine Art Wahrscheinlichkeitsmaß[1] betrachtet wird, dass ein Nutzer beim Durchstöbern des Forums auf diese Nachricht stößt, nimmt im Lauf der Zeit automatisch ab, während Verweise anderer Nachrichten auf diese Nachricht die Sichtbarkeit wieder erhöhen. Es bestehen verschiedene Möglichkeiten, aus diesem allgemeinen Zusammenhang ein Sichtbarkeitsmaß zu entwickeln. Malsch und Schlieder schlagen zunächst die folgende Funktion vor:

$$\text{msgvisibility}(m) = \sum_{n \in \{m\} \cup DP_m} e^{-t(n)}$$

Hierbei ist m die betrachtete Nachricht, DP_m die Menge aller direkten Vorgänger von m, d.h. alle Nachrichten, die direkt auf m verweisen. t(n) ist eine Zeitfunktion, die beginnend bei 0

[1] Es sind keine wirklichen Wahrscheinlichkeiten – schon auf Grund des möglichen Wertebereichs von Sichtbarkeiten – doch die Vorstellung ist an dieser Stelle hilfreich.

für die Gegenwart jeder Nachricht eine positive ganze Zahl als Timestamp zuweist. In Abbildung 5 würde der Nachrichtengeneration {8, 9, 10, 11}der Timestamp 0 zugewiesen werden, der Generation {6, 7} eine 1 usw.
Abbildung 6 zeigt die mit dieser Formel ermittelten Sichtbarkeiten für die Nachrichten aus Abbildung 5:

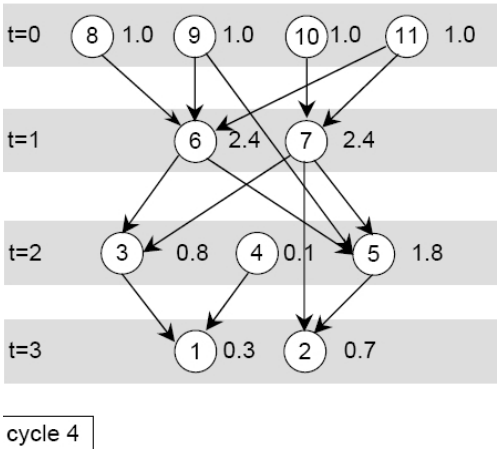


Abbildung 6: Sichtbarkeiten der Nachrichten aus Abbildung 5 (vgl. [MaSchlie02], Fig. 4)

2.2.2.2 Sichtbarkeitsmaß für Themen in Diskussionsforen

Die in 2.1 beschriebene Definition lässt sich nun wie folgt spezialisieren:

DOCUMENTS ist in diesem Fall eine Menge von Nachrichten.

Eine Selektion der Nachrichten nach Threads innerhalb eines Forums oder der Vergleich mehrerer Foren untereinander sind hier als Analyse verschiedener Aktoren denkbar.

relevance werde mit Hilfe von Methoden des Information Retrieval für alle Nachrichten beliebig definiert.

structure ist ein gerichteter azyklischer Graph zur Abbildung der Verweisstruktur zwischen den Nachrichten. Der Verfasser einer neuen Nachricht im Diskussionsforum verknüpft seine erstellte Nachricht meist mit einer (Baum) oder mehreren (Graph) bereits existierenden Nachrichten. Weiterhin ist jeder Nachricht ein Zeitpunkt ihrer Entstehung zuge-

ordnet, d.h. es existieren mehrere Generationen von Nachrichten (vgl. Abbildung 3 bzw. Abbildung 5)

msgvisibility(msg) bezeichnet die Sichtbarkeit einer Nachricht (msg∈DOCUMENTS), wie sie sich rein aus der gegebenen Verweisstruktur errechnen lässt (vgl. 2.2.2.1)

visibility(top) errechnet sich dann aus der Sichtbarkeit der Nachrichten und der Relevanz dieser Nachrichten für ein bestimmtes Thema z.B. wie folgt:

$$\text{visibility(top)} = \sum_{\text{msg}\in\text{DOCUMENTS}} \text{relevance}(\text{top}, \text{msg}) * \text{msgvisibility(msg)}$$

Der wesentliche Unterschied bei der Ermittlung von Sichtbarkeitswerten bei den Internetforen zu der bei Webseiten liegt also – neben der unterschiedlichen technischen Realisierung – darin, dass die Strukturinformation auf keinen Fall vernachlässigt werden sollte. Grund ist, dass in einem Internetforum im Allgemeinen tatsächlich entlang von Verweisen navigiert wird: Der User hangelt sich von Nachricht zu Nachricht, während bei Webseiten sehr häufig auch ein „Quereinstieg“ über einen Direktlink einer Suchmaschine erfolgt. Ein weiterer Grund ist, dass im Forum durch die in der Struktur enthaltene Zuordnung von Zeitpunkten zu Nachrichten ein Veralten von Nachrichten einrechenbar ist, während ein Timestamp für Webpages fehlt. Ein weiterer zu erwähnender Unterschied ist, dass Foren (bzw. Teilforen) oft auf ein bestimmtes Themengebiet eingeschränkt sind, was einen effizienteren Einsatz des Information Retrieval für die relevance-Funktion ermöglicht.

2.3 Co-Sichtbarkeit

Die oben erwähnte Problemstellung, ob ein Thema aus einem anderen hervorgeht oder zwei Themen tendenziell gemeinsam diskutiert werden, ist im Rahmen der Sichtbarkeitsanalyse auf die Frage zurückzuführen, ob zwei Themen gemeinsam sichtbar sind[1]. Die Formulierung „gemeinsam“ lässt sich hierbei auf zwei verschiedene Arten auffassen.

[1] Die nun folgenden Ausführungen lassen sich selbstverständlich auch auf die gemeinsame Sichtbarkeit von n Themen (n >= 2) verallgemeinern. Darauf wird an dieser Stelle verzichtet, da dies nicht zum tieferen Verständnis

2.3.1 Co-Sichtbarkeit zum gleichen Zeitpunkt

Zwei Themen haben genau dann eine hohe gemeinsame Sichtbarkeit, wenn ihre (Einzel)Sichtbarkeiten gleichzeitig hoch sind. Die Co-Sichtbarkeit hängt in diesem Fall nur von den zwei Sichtbarkeitswerten der Themen ab und die allgemeinen Definitionen für Sichtbarkeitsmaße aus 2.1 können folgendermaßen ergänzt werden:

covisibility: $IR^{+}_{0} \times IR^{+}_{0} \rightarrow [0..1]$

bezeichnet die Co-Sichtbarkeitsfunktion, die je zwei Themen top_1, $top_2 \in TOP$ auf Grund ihrer visibility-Werte visibility(top_1) und visibility(top_2) einen Co-Sichtbarkeitswert im Intervall [0..1] zuordnet.

Diese Auffassung von Co-Sichtbarkeit greift jedoch in vielen Fällen zu kurz, denn sie berechnet lediglich, wann zwei Themen gleichzeitig auftreten, nicht jedoch, wo sie zur gleichen Zeit auftreten. So könnte beispielsweise zwar in einem Thread eines Forums intensiv über den verstorbenen Rudolf Mooshammer diskutiert werden und zum gleichen Zeitpunkt in einem anderen davon unabhängigen Thread über den genetischen Fingerabdruck, was jedoch nicht bedeuten würde, dass diese beiden Themen notwendigerweise in einem Zusammenhang stünden. Für derartige Folgerungen ist eine restriktivere Sicht der Co-Sichtbarkeit notwendig.

2.3.2 Co-Sichtbarkeit in den gleichen Dokumenten (und zum gleichen Zeitpunkt)

Zwei Themen haben genau dann eine hohe gemeinsame Sichtbarkeit, wenn es viele Dokumente gibt, die zum gleichen Zeitpunkt eine hohe Relevanz für beide Themen besitzen. Im Unterschied zu der in 2.3.1 erläuterten Definition der Co-Sichtbarkeit wird hier also nicht nur der gleiche Zeitpunkt, sondern noch das gemeinsame Auftreten in denselben Dokumenten gefordert.

---

der Co-Sichtbarkeit beiträgt und eine solche Verallgemeinerung in den folgenden Abschnitten nicht benötigt wird.

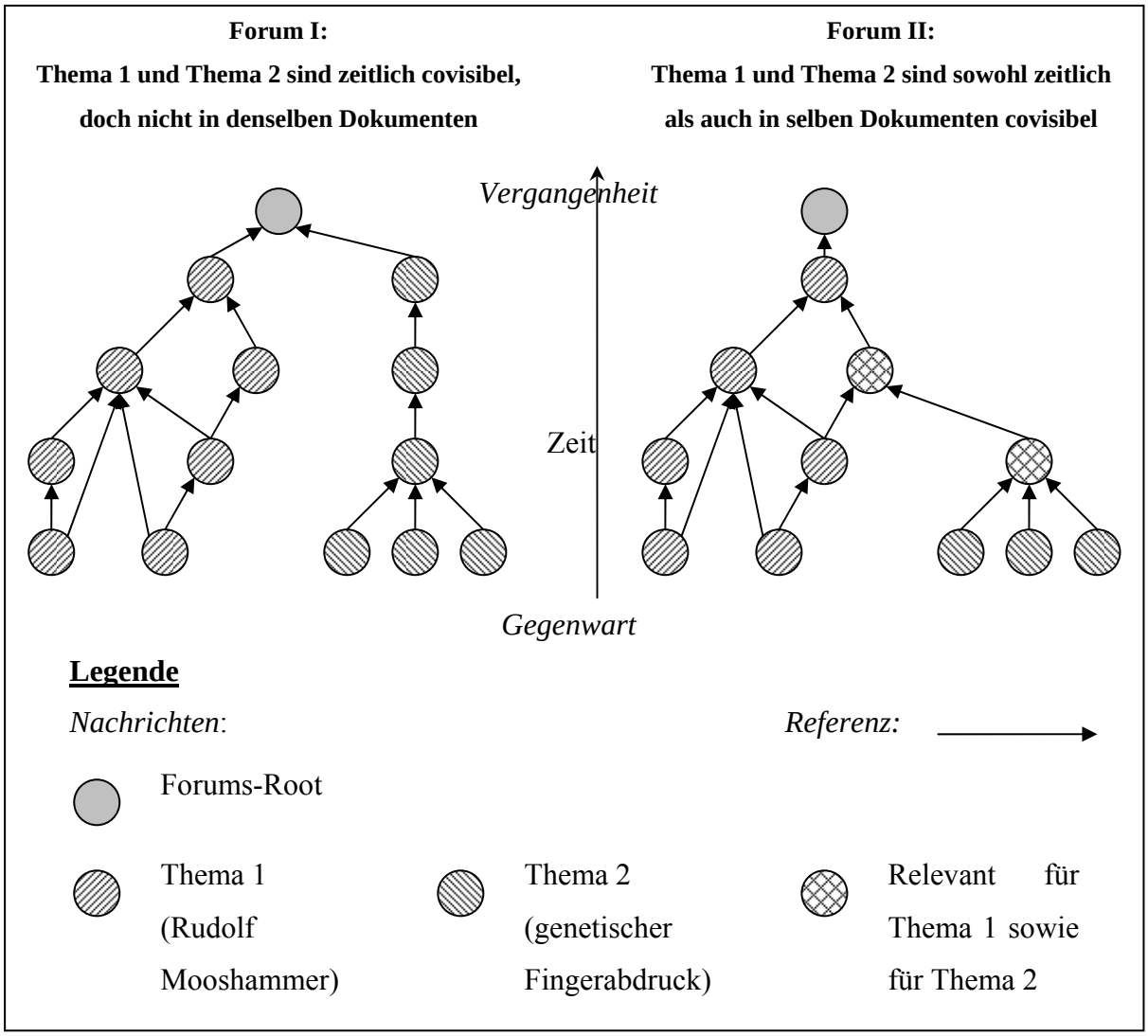


Abbildung 7: Referenzstruktur zwischen Nachrichten mit zugeordneten Themen in zwei verschiedenen Internetforen als Beispiel für verschiedene Arten von Co-Sichtbarkeit

Abbildung 7 zeigt den Unterschied der beiden Auffassungen von Co-Sichtbarkeit am Beispiel eines Diskussionsforums: Während in Forum I in zwei getrennten Threads unabhängig voneinander über Thema 1 und Thema 2 diskutiert wird, entwickelt sich die Diskussion in Forum II von Thema 1 zu Thema 2 – der Übergang wird an den zwei Nachrichten deutlich, die für beide Themen relevant sind.

Die Definition einer Co-Sichtbarkeit sähe für diese zweite Auffassung von Co-Sichtbarkeit folgendermaßen aus:

corelevance: TOP × TOP × DOCUMENTS → [0..1]

ist eine Relevanzfunktion, die den Dokumenten einen Relevanzwert für je zwei Themen zuordnet. Anmerkung: Die Definition eines corelevance-Maßes unabhängig von den relevance-Maßen der einzelnen Themen ermöglicht es, Einzelheiten des Auftretens der Themen innerhalb des Dokuments zu berücksichtigen, z.B. die Nähe der Begriffe zueinander. Wie auch bei der relevance-Funktion in 2.1 sei hier auf die Standardliteratur des Information Retrieval verwiesen (vgl. z.B. [BaezaRib04]).

covisibility$_{\text{DOCUMENTS, corelevance, structure}}$**:** TOP × TOP → [0..1]

ist dann entsprechend zur visibility-Funktion die Co-Sichtbarkeitsfunktion unter Berücksichtigung der Co-Relevanz

2.3.2.1 Co-Sichtbarkeitsmaß für Themen auf Webseiten

Auf eine Konkretisierung der allgemeinen Co-Sichtbarkeits-Definition für Diskussionsforen wird an dieser Stelle verzichtet, doch eine (abermals sehr einfache und mit Recall-Werten einer Suchmaschine arbeitende) Konkretisierung für Themen auf Webseiten soll im Folgenden als Ergänzung zu 2.2.1.1 vorgeschlagen werden.

corelevance

corelevance(top_1, top_2, doc) = min(relevance(top_1, doc), relevance(top_2, doc)),

d.h. 1, falls beide Zeichenketten enthalten sind, ansonsten 0.

Aufsummiert über alle Dokumente entspricht dies in unserem einfachen Fall (vgl. 2.2.1.1) der Trefferzahl einer Suchmaschine auf die Anfrage „top_1 AND top_2“

covisibility

$$\text{covisibility}_1(top_1, top_2) = \frac{\sum_{doc \in DOCUMENTS} \text{corelevance}(top_1, top_2, doc)}{\text{visibility}(top_1)}$$

$$\text{covisibility}_2(top_1, top_2) = \frac{\sum_{doc \in DOCUMENTS} \text{corelevance}(top_1, top_2, doc)}{\text{visibility}(top_2)}$$

Die Co-Sichtbarkeit ist hier zunächst der Anteil der Seiten, die beide Themen enthalten, an den Seiten, die eines der beiden Themen enthalten. Daraus ergeben sich zwei Co-Sichtbarkeitsmaße covisibility_1 und covisibility_2. Eine weitergehende Interpretation der beiden möglichen Werte bleibt schwierig, doch die Unterscheidung zweier eigener Maßzahlen ist notwendig, da im weiteren Verlauf dieser Arbeit einer von ihnen besonderen Einsatz finden wird.

Eine Definition, die nur zu einem möglichen Wert führt, ergibt sich aus dem Anteil der Seiten, die beide Themen enthalten, an den Seiten, die mindestens eines der beiden Themen enthalten.

$$\text{covisibility}(top_1, top_2) = \frac{\sum_{doc \in DOCUMENTS} \text{corelevance}(top_1, top_2, doc)}{\text{visibility}(top_1) + \text{visibility}(top_2) - \sum_{doc \in DOCUMENTS} \text{corelevance}(top_1, top_2, doc)}$$

Tabelle 1 zeigt die (geschätzten) Trefferzahlen der Suchmaschine Google™ für die Themen HIV und Aids sowie die daraus errechneten Co-Visibilitäten. Abbildung 8 visualisiert die einzelnen und gemeinsamen Trefferzahlen der beiden Themen in einem Balken.

Trefferzahl HIV	6.410.000
Trefferzahl Aids	10.000.000
Trefferzahl HIV AND Aids	2.550.000
covisibility(HIV, Aids)	0,183983
covisibility_1(HIV, Aids)	0,397816
covisibility_2(HIV, Aids)	0,255000

Tabelle 1: Trefferzahlen (der Suchmaschine Google™) und Co-Visibilitäten der Themen „Aids“ und „HIV“

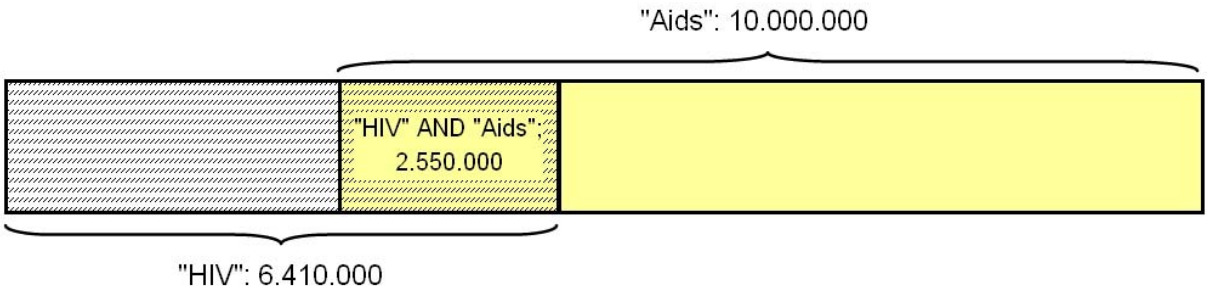


Abbildung 8: Trefferzahlen (der Suchmaschine Google™) von „Aids“ und „HIV“ sowie „Aids AND HIV“

3. PROBLEMSTELLUNG: BERÜCKSICHTIGUNG SEMANTISCHER BEZIEHUNGEN ZWISCHEN THEMEN

Die genaue Zielsetzung dieser Arbeit wird im Folgenden an einem ausführlichen Beispiel illustriert.

Ein Anwender der Sichtbarkeitsanalyse interessiere sich für folgende Themen:

{HIV, Aids, Geschlechtskrankheit, Safer Sex, Syphilis, Kondom, Präservativ}

Unter Anwendung der in 2.2.1 beschriebenen Methoden zu Ermittlung der Sichtbarkeit von Themen auf Webseiten mit Hilfe der Trefferzahl einer Suchmaschine erhalte er für eine bestimmte Domain die in Tabelle 2 gelisteten Sichtbarkeitsdaten, im Folgenden „initiale Sichtbarkeiten“ genannt:

Thema	(initiale) Sichtbarkeit
HIV	10
Aids	100
Geschlechtskrankheit	5
Safer Sex	5
Syphilis	20
Kondom	30
Präservativ	0

Tabelle 2: Initiale Sichtbarkeiten verschiedener Themen aus dem Themengebiet „Safer Sex“

Abbildung 9 und Abbildung 10 visualisieren die Verhältnisse der initialen Sichtbarkeiten zueinander in einem Säulen- bzw. Kreisdiagramm. Der prozentuale Anteil einer Sichtbarkeit an der Summe aller Sichtbarkeiten wird in dieser Arbeit relative Sichtbarkeit genannt, hier also initiale relative Sichtbarkeit.

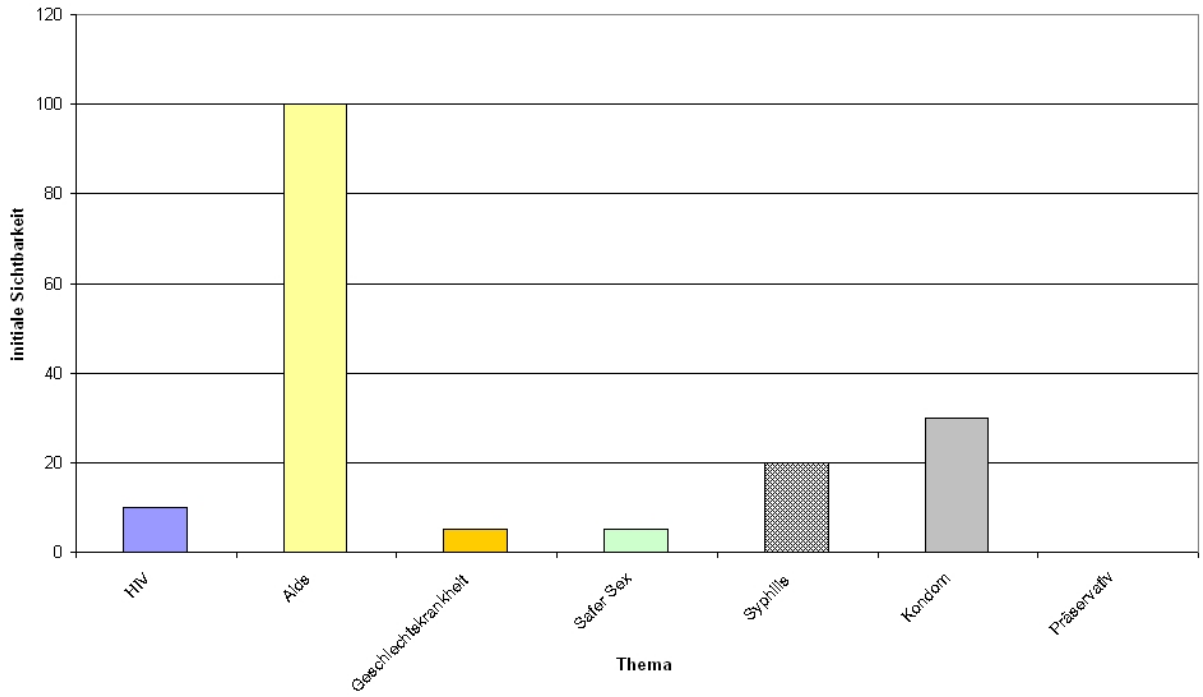


Abbildung 9: Initiale Sichtbarkeiten verschiedener Themen aus dem Themengebiet „Safer Sex“

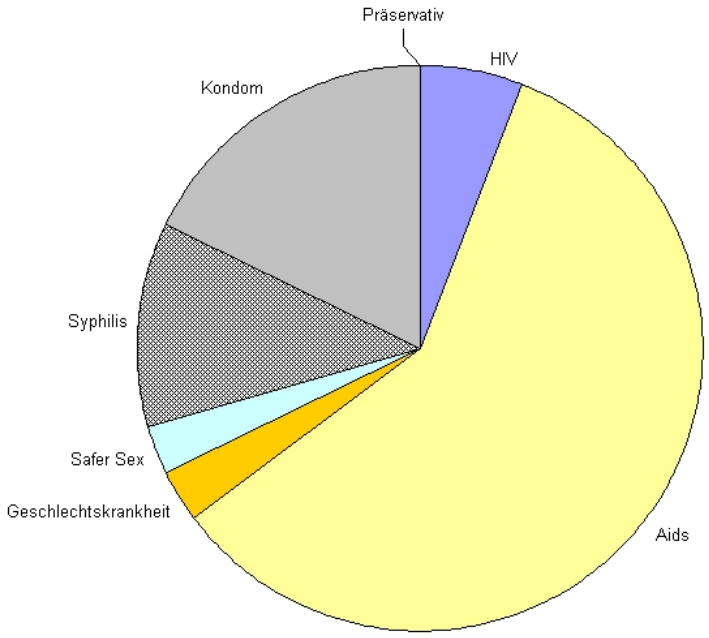


Abbildung 10: Initiale relative Sichtbarkeiten verschiedener Themen aus dem Themengebiet „Safer Sex“

Versucht man nun, diese Daten zu interpretieren, so stößt man auf offensichtliche Probleme: Die Frage „Wie hoch ist die Sichtbarkeit des Themas ‚Präservativ'?“ muss beantwortet werden mit: „Dieses Thema verfügt über gar keine Sichtbarkeit.“ Das heißt, die vorliegenden Daten suggerieren, dass dieses Thema in der beobachteten Domain überhaupt nicht vor-

kommt, was jedoch nicht den Tatsachen entspricht, denn Präservativ ist schließlich nichts anderes als der Fachausdruck für Kondom und das Thema „Kondom“ ist durchaus sichtbar. Der Grund für die Null-Visibilität von „Präservativ“ liegt also nicht an mangelndem Interesse an diesem Thema, sondern wohl daran, dass der in der Umgangssprache häufiger vorkommende Begriff „Kondom“ verwendet wurde.
Ähnliche Probleme ergeben sich bei der Frage nach der Sichtbarkeit von „Geschlechtskrankheit“: Es erscheint auf den ersten Blick sehr wenig sichtbar, doch die Themen „HIV“ und „Syphilis“ sind beispielsweise sicherlich mit „Geschlechtskrankheit“ verwandt. Eine Diskussion oder Abhandlung zum Thema „HIV“ handelt immer ganz automatisch auch über eine Geschlechtskrankheit. Anders gesagt: Jede Webseite, auf der das Thema „HIV“ sichtbar ist, erhöht auch gleichzeitig die Sichtbarkeit des Themas „Geschlechtskrankheit“.

Allgemein formuliert wurden bisher die semantischen Beziehungen zwischen den Themen nicht berücksichtigt. Fokus dieser Arbeit soll es sein, einen Ansatz für dieses Problem zu entwickeln, der einen Sichtbarkeitsausgleich durchführt und aus den bisherigen unrealistischen initialen Sichtbarkeitswerten geeignetere Sichtbarkeiten ableitet.

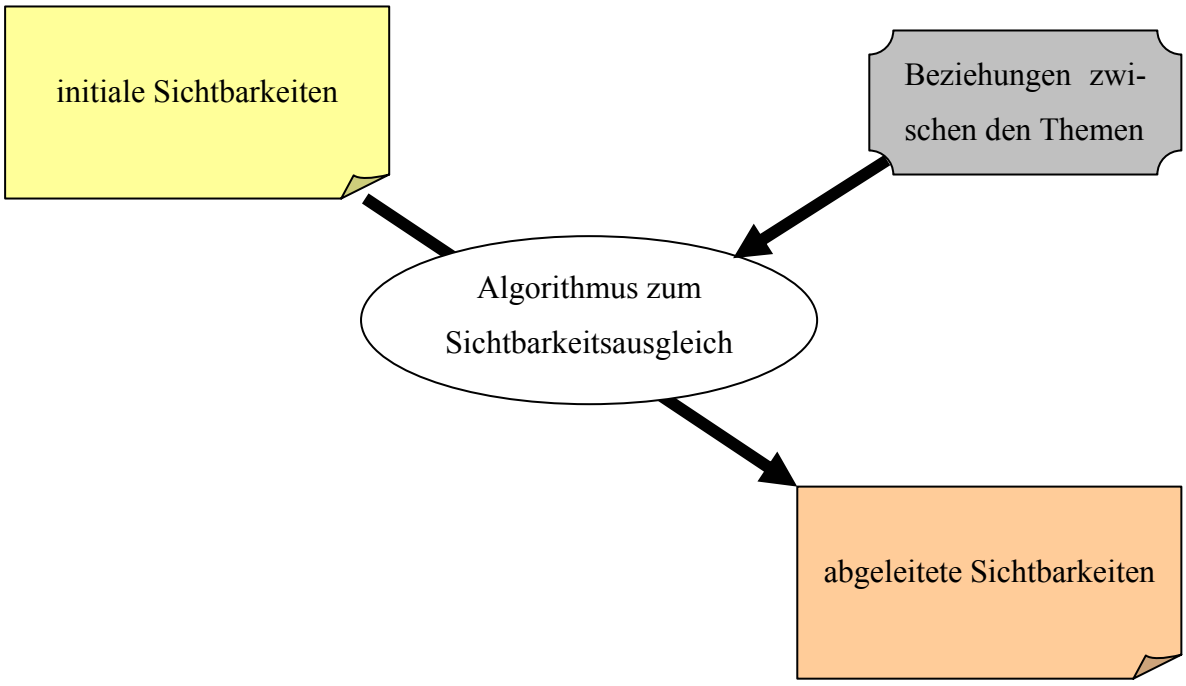


Abbildung 11: Sichtbarkeitsausgleich

3.1 Modellierung der semantischen Beziehungen zwischen Themen

Das hierbei als erstes zu lösende Teilproblem ist, die semantischen Beziehungen zwischen den Themen zu modellieren.

Formal äußert sich das in der Ergänzung der Definition aus 2.1:

topicrelationships

bezeichnet Information über die Beziehungen zwischen den Themen aus TOP. Es werden Beziehungen zwischen je zwei Themen top_1, $top_2 \in$ TOP abgebildet, die den folgenden Satz erfüllen:

„Wenn in einem Dokument top_1 sichtbar ist, so ist in diesem Dokument (zu einem gewissen Grad/mit einer gewissen Wahrscheinlichkeit) auch top_2 sichtbar."

Anforderungen an eine solche Modellierung sind:

- Es muss möglich sein, gerichtete Beziehungen zu modellieren.

  Es ist denkbar, dass zwar ein Thema 1 enthaltendes Dokument die Sichtbarkeit von Thema 2 erhöht, jedoch nicht umgekehrt. Im oben genannten Beispiel wäre das zwischen „Syphilis" und „Geschlechtskrankheit" der Fall: Ein Dokument, das von der Syphilis handelt, thematisiert gleichzeitig automatisch eine Geschlechtskrankheit (auch wenn der String „Geschlechtskrankheit" selbst gar nicht vorkommt). Hingegen ist es nicht so, dass Dokumente über Geschlechtskrankheiten stets auch das Thema „Syphilis" enthalten.

- Es muss möglich sein, die Stärke einer Beziehung auszudrücken.

  Beispielsweise besteht mit Sicherheit eine sehr starke Beziehung von „HIV" in Richtung „Aids", da das HI-Virus später das Krankheitsbild Aids verursacht und ohne das HI-Virus kein Aids denkbar ist. Eine etwas schwächere Beziehung hingegen besteht zwischen „Kondom" und „Safer Sex": Nur Seiten, die Kondome aus der Sicht der Sexualaufklärung behandeln, sind hier mitzuzählen, nicht jedoch jene, die Kondome vermarkten oder etwa als Scherzartikel darstellen.

- Die Modellierung sollte für den Nutzer möglichst intuitiv handhabbar sein und komplexe Beziehungen unterstützen.

  So ist die Möglichkeit von Synonymen wünschenswert (im Beispiel: „Kondom“ und „Präservativ“), da diese Art von Beziehung sicher anders gehandhabt werden muss als eine einfache Assoziation („Safer Sex“ → „Geschlechtskrankheit“). Eine weitere essentielle Beziehungsart sind Hyponymien – im Beispiel: „Syphilis“ is_a „Geschlechtskrankheit“.

- Da die modellierten Beziehungen ein Input für den Algorithmus sind, sollte der Modellierungs-Output aus Gründen der Komplexitätsbewältigung nicht allzu kompliziert sein.

  Ein beliebiges Kombinieren einer großen Anzahl von Beziehungstypen macht es unter Umständen schwierig, die Korrektheit und Angemessenheit eines auf die Beziehungen anzuwendenden Algorithmus' darzulegen.

3.2 Algorithmischer Ausgleich der Sichtbarkeiten von Themen

Gegeben die semantischen Beziehungen zwischen den Themen ist im nächsten Schritt das eigentliche Hauptproblem zu lösen: Die Durchführung des Sichtbarkeitsausgleichs durch einen Algorithmus. Ungeachtet der Tatsache, dass sich die exakten Anforderungen an den Algorithmus erst nach Festlegung der Art der semantischen Beziehungen aufstellen lassen, können folgende allgemeinen Forderungen gestellt werden:

- Keine Abnahme von Sichtbarkeiten:

  Die Sichtbarkeit eines Themas soll durch den Algorithmus nicht abnehmen können. Eine bereits vorhandene Sichtbarkeit kann nicht verschwinden. Begründung hierfür ist das Beispiel der Webseiten: Fasst man die Sichtbarkeitswerte als Anzahl relevanter Seiten auf, sollen durch den Algorithmus keine Seiten abgezogen werden, sondern lediglich einige Seiten bei anderen Themen „mitgezählt“ werden.

  Das Verhältnis der Sichtbarkeiten der Themen zueinander kann sich hingegen verändern, so dass eine Abnahme der relativen Sichtbarkeit möglich ist.

- Unterstützung transitiver Beziehungen:

  Im Beispiel soll eine Beziehung zwischen „Kondom“ und „Safer Sex“ sowie eine Beziehung von „Safer Sex“ nach „Geschlechtskrankheiten“ sich beim Sichtbarkeitsausgleich so auswirken, dass die Wirkung von „Kondom“ transitiv auch in gewissem Grad auf „Geschlechtskrankheit“ ausstrahlt.

- Eine kleine Veränderung der initialen Werte darf sich nur in einer kleinen Veränderung der algorithmisch ermittelten Endwerte niederschlagen.

  So sollte ein initialer Sichtbarkeitswert von 101 für „Aids“ keinen fundamentalen Unterschied zu einer initialen Sichtbarkeit von 100 machen.

- Eine kleine Veränderung der Stärke einer Beziehung darf sich nur in einer kleinen Veränderung der Endwerte zeigen.

- Im Idealfall ist eine Berücksichtigung der Covisibilitäten wünschenswert.

- Der Algorithmus muss „sinnvolle“ Werte liefern.

  Diese Forderung ist schwammig und kann nur intuitiv überprüft werden. Beispielsweise wäre es sicher nicht sinnvoll, wenn die Auswirkung der Sichtbarkeit von „Aids“ (initial 100) auf HIV (initial 10) so groß wäre, dass am Ende Aids die Sichtbarkeit von 100 und HIV die Sichtbarkeit von 100.000 zugeordnet wird.

4. MODELLIERUNG DER SEMANTISCHEN BEZIEHUNGEN ZWISCHEN THEMEN

4.1 Themennetzwerk als gerichteter, gewichteter Graph

Eine erste sehr einfache Idee, die semantischen Beziehungen zwischen den Themen darzustellen, besteht in der Verwendung eines gerichteten und gewichteten Graphen. Jedes Thema wird auf einen Knoten abgebildet, wobei der Knoten mit der Sichtbarkeit des Themas gewichtet wird (am Anfang die initiale Sichtbarkeit). Eine Kante von Knoten top_1 nach top_2 bedeutet, dass zwischen diesen beiden Themen eine semantische Beziehung derart besteht, dass eine hohe Sichtbarkeit von top_1 auch die Sichtbarkeit von top_2 erhöht[1]. Die Stärke der semantischen Beziehung wird durch eine Kantengewichtung im Intervall von 0 (keine Beziehung) bis 1 (starke Beziehung) ausgedrückt. Ein solcher gerichteter und gewichteter Graph aus Themen soll Themennetzwerk genannt werden.

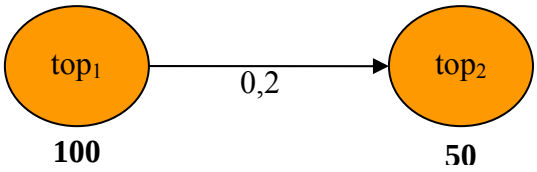


Abbildung 12: Themennetzwerk mit zwei Knoten und einer Kante

topicrelationships (aus 3.1) ist hier also formal gesehen eine Abbildung von je zwei Themen auf eine Zahl im Intervall [0..1], die das Kantengewicht ausdrückt (oder keine Beziehung bei 0).

[1] In Abschnitt 3.1 wurde dies am Beispiel der Webseiten mit folgendem Satz beschrieben: „Wenn in einem Dokument top_1 sichtbar ist, so ist in diesem Dokument (zu einem gewissen Grad/mit einer gewissen Wahrscheinlichkeit) auch top_2 sichtbar.“

$$\text{topicrelationships: TOP} \times \text{TOP} \rightarrow [0..1]$$

Auf die Möglichkeit negativer Kantengewichte wird bewusst verzichtet. Der Grund hierfür wurde bereits in 3.2 bei den Anforderungen an den Algorithmus beschrieben: Die Abnahme von Sichtbarkeit – und nichts anderes würde ein negatives Kantengewicht bedeuten – wird ausgeschlossen. Inhibition zwischen zwei Themen (Sichtbarkeit von Thema 1 hemmt die Sichtbarkeit von Thema 2) ist in dem hier beschriebenen Ansatz, der von Webdokumenten ausgeht, nicht sinnvoll interpretierbar: Die Sichtbarkeit eines Themas, die durch real existierende Dokumente geschaffen wird, kann nicht vermindert, sondern lediglich durch andere Dokumente mit semantisch verwandten Themen erhöht werden.

Dieser erste einfache Ansatz erfüllt bereits viele der in 3.1 formulierten Anforderungen an die Modellierung der Themen: Die Kanten sind gerichtet und die Stärke der Beziehung wird durch eine Kantengewichtung ausgedrückt. Weiterhin ermöglicht die Verwendung eines Graphen die volle Anwendung der Methoden der Graphentheorie und Netzwerkanalyse (vgl. z.B. [Diestel00], [Hanneman01], [Jansen99]). Die Anzahl der möglichen Beziehungstypen zwischen Themen beschränkt sich auf eine; alle denkbaren Beziehungstypen (z.B. is_a, Assoziation, Synonym) werden durch die gleiche Kantenart ausgedrückt, so dass der Modellierungsoutput aus Sicht des folgenden Algorithmus' wenig komplex ist. Doch dies schränkt auf der anderen Seite den Nutzen aus Sicht des Modellierers stark ein, der sich bei der Modellierung differenzierte Beziehungsarten wünscht. Gerade falls ein solches Themennetzwerk nach einiger Zeit verändert werden soll, ist es unter Umständen nicht mehr auf Anhieb möglich, die genaue Semantik der Kanten nachzuvollziehen. Aus diesem Grund wird in 4.2 auf Modellierungsansätze mit größeren semantischen Ausdrucksmöglichkeiten eingegangen.

Abbildung 13 zeigt ein Beispiel für ein etwas größeres Themennetzwerk mit zyklischen sowie transitiven Beziehungen. Die Kantengewichte sind in diesem Fall sehr hoch gewählt, da die hier verwendeten Themen stark miteinander verwandt sind. Selbstverständlich ist jedes Modell stets aus Sicht des Modellierers erstellt und spiegelt dessen Sicht auf den modellierten Teilausschnitt der Diskurswelt dar (siehe auch: konstruktivistisches Weltbild der Modellierung, vgl. z.B. [Goorhuis94]). Daher lässt sich über jede Modellierung diskutieren (beispielsweise über die hier gewählte Höhe der Kantengewichte), was jedoch Aufgabe der Anwender

einer Sichtbarkeitsanalyse ist. Dem Leser soll demnach an dieser und anderen Stellen dieser Arbeit eine detaillierte Begründung für die gewählte Modellierung erspart bleiben.

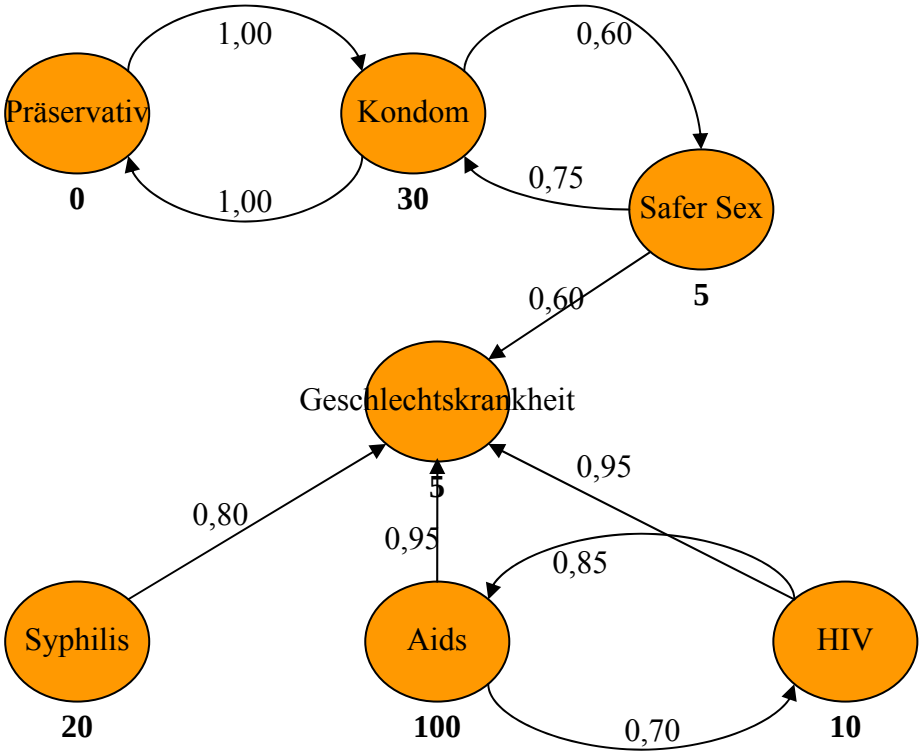


Abbildung 13: Themennetzwerk zum Themengebiet „Safer Sex“

4.2 Modellierungstechniken aus der relevanten Fachliteratur

4.2.1 Semantische Netze

Semantische Netze dürften sicherlich bekannt sein und werden daher nur kurz umschrieben (für eine umfassende Behandlung Semantischer Netze vgl. z.B. [Sowa00]).

Ein Semantisches Netz besteht aus einer Menge von Knoten, die durch Bögen verbunden werden. Knoten und Bögen werden jeweils mit Namen bezeichnet. Knoten repräsentieren Objekte (Instanzen, Konzepte, Begriffe), zwischen denen Relationen (beschrieben durch die Bögen) bestehen. Semantische Netze basieren auf kognitionspsychologischen Modellen des menschlichen Gedächtnisses und sind grundsätzlich zur Beschreibung von verschiedensten Relationen verwendbar, so dass sich jede Art von Wissen repräsentieren lässt.

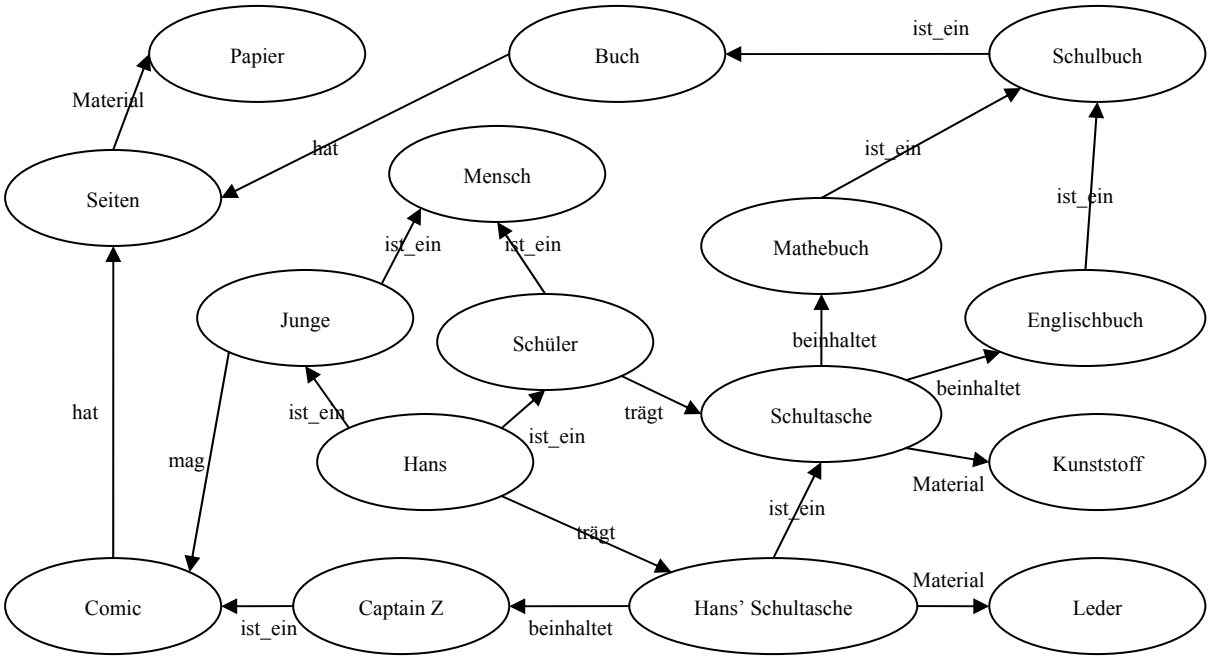


Abbildung 14: Beispiel für Wissensrepräsentation in einem Semantischen Netz

Abbildung 14 zeigt ein Beispiel für ein allgemeines Semantisches Netz. Verschiedenste Arten von Beziehungen (ist_ein, hat, trägt, Material, …) bestehen zwischen den Objekten, wobei allgemeine Konzepte (Schüler, Buch, Mensch, Schultasche) und Instanzen (Hans, Hans' Schultasche) nebeneinander verwendet werden. Typische objektorientierte Ideen wie Vererbung oder Überschreiben von Eigenschaften sind hier entsprechend anwendbar.

Für die Modellierung der Themen für die Sichtbarkeitsanalyse stellen Semantische Netze (im Gegensatz zu den Themennetzwerken aus 4.1) alle denkbaren Beziehungstypen zur Verfügung, was dem Nutzer große Freiräume verschafft. Diese beliebigen Beziehungstypen sind jedoch ein immenses Komplexitätsproblem für den folgenden Algorithmus, der nun nicht nur eine, zwei oder drei Beziehungsarten verarbeiten muss, sondern auch mit bisher unbekannten Beziehungen sinnvoll umgehen sollte. Weiterhin wäre die Stärke der Beziehung zwischen zwei Themen nur schwer auszudrücken, da Bögen in Semantischen Netzen lediglich beschriftet sind – nicht jedoch gewichtet. Dies ließe sich beispielsweise durch den in Abbildung 15 dargestellten „Workaround“ umgehen.

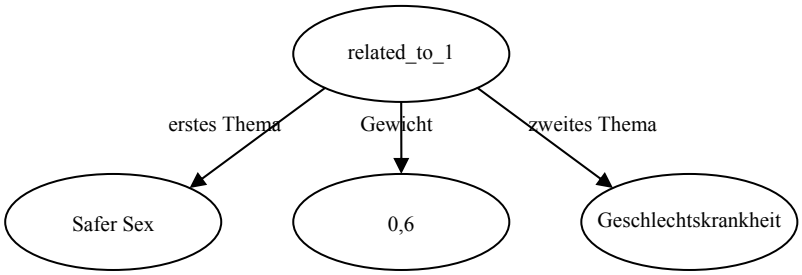


Abbildung 15: Gewichtete related_to-Beziehung im Semantischen Netz

In großen Netzen wäre dies mit Sicherheit unübersichtlich. Außerdem wird implizit unterstellt, dass der Nutzer Beziehungsknoten vom Typ „related_to“[1] stets mit genau zwei Konzeptknoten und einem Gewichtsknoten verknüpft, wobei die Bögen mit „erstes Thema“, „zweites Thema“ bzw. „Gewicht“ benannt werden und der Gewichtsknoten eine Zahl im Intervall [0…1] enthält.

Die genannten Probleme allgemeiner Semantischer Netze machen es wünschenswert, die Arten der möglichen Beziehungen für die Modellierung von Themen zu beschränken und zu standardisieren, was durch die im Folgenden beschriebenen Ansätze verfolgt wird.

4.2.2 Topic Maps

Topic Maps (vgl. z.B. [MüWid02], [Pepper00]) sind eine international standardisierte Art der Wissensrepräsentation (Standards: [ISO13250] auf dem Datenformat SGML und HyTime beruhend; bzw. [XTM1.0] XML-basiert), die ebenfalls beliebige Beziehungstypen zulässt. Obwohl sie auf den ersten Blick ähnlich zu Semantischen Netzen sind, weisen Topic Maps Konzepte auf, die einen weitaus systematischeren Einsatz erlauben:

- Topics[2]: Ein Topic entspricht dem Knoten im Semantischen Netz und kann wie dort für ein beliebiges Objekt stehen. Jedes Topic kann einen oder mehrere Typen haben, wobei jeder Typ selbst wiederum als Topic modelliert wird. Die Möglichkeit zu Typhierarchien

[1] Im Beispiel wurde davon ausgegangen, dass in einem Semantischen Netz keine zwei Knoten mit der gleichen Beschriftung vorkommen dürfen. Daher „related_to_1“ als eine konkrete Beziehung vom Typ „related_to“. Entsprechend denkbar ist z.B. „is_a_1“ als eine konkrete Beziehung vom Typ „is_a“.

[2] Das Konzept „Topic“ aus der Topic Map sei an dieser Stelle nicht zu verwechseln mit den in 2.1 eingeführten Themen, die in dieser Arbeit auch oft als Topic oder top_i bezeichnet werden. Ein Topic aus der Topic Map kann zwar für ein Thema stehen, grundsätzlich jedoch auch für jedes andere beliebige Objekt.

wie in Abbildung 14 (Hans → Junge → Mensch) ist also im Standard explizit enthalten. Solche Hierarchien wären auch in der Modellierung von Themen einsetzbar (HIV → Geschlechtskrankheit → Infektionskrankheit → Krankheit).

- Public Subject Descriptor: Topic Maps erlauben es, einem Topic über einen Id-Wert eine Identität zuzuweisen. Nützlich ist dies meist beim Zusammenführen mehrerer Topic Maps, in denen identische Topics mit verschiedenem Namen auftreten (z.B. verschiedene Sprachen: „Deutschland“, „Germany“). Themen könnten über den Public Subject Descriptor als identisch gekennzeichnet werden (z.B. „Kondom“ = „Präservativ“), so dass eine Beziehung vom Typ „Synonym“ nicht mehr notwendig ist.

- Occurrences: Jedem Topic sind beliebig viele Occurrences zugeordnet, die Verweise auf externe Informationsquellen darstellen. So könnte ein Topic „Helmut Kohl“ auf ein Webdokument mit dessen Lebenslauf oder ein Photo verweisen. Für die Modellierung der semantischen Beziehungen zwischen Themen birgt dies keinen wesentlichen Vorteil, da solche Links nur für den Benutzer Zusatzinformation liefern und keinerlei Einfluss auf den Algorithmus zum Sichtbarkeitsausgleich hätten.

- Scopes erlauben ein Zusammenfassen von Topics zu einem Gültigkeitsbereich, wodurch der Umgang mit Homonymen möglich wird: Zwei Topics dürfen den gleichen Namen haben, solange sie sich in unterschiedlichen Gültigkeitsbereichen befinden. In 2.1 wurde bereits erwähnt, dass Themen in dieser Arbeit eh durch ihre Namen identifiziert werden, so dass Scopes hier keinen Nutzen bringen.

- Associations in Topic Maps stellen eine Erweiterung der Bögen in Semantischen Netzen dar: Im Standard ist die Möglichkeit mehrstelliger Relationen vorgesehen, d.h. eine Association kann beliebig viele Topics verbinden. Eine Association kann von einem bestimmten Association Type sein und jedem an einer Association teilnehmendem Topic kann eine Rolle (Association Role) zugewiesen werden. Association Types und Association Roles werden selbst als Topics in der Topic Map modelliert, wodurch eine Formalisierung der Beziehungstypen möglich ist – ein wesentlicher Vorteil zu den Semantischen Netzen. Verwendet man bei der Themenmodellierung nun ausschließlich typisierte Associations,

so sind der Beliebigkeit beim Einsatz von Beziehungstypen insofern Grenzen gesetzt, dass jeder Beziehungstyp zunächst in der Topic Map definiert werden muss.

- Facets ermöglichen die Zuordnung von beliebigen Eigenschaftswerten zu Topics oder Associations. Eine gewichtete Beziehung ließe sich dadurch bei weitem intuitiver und übersichtlicher modellieren als beim Semantischen Netz (vgl. Abbildung 15 versus Abbildung 16): Eine Association vom Typ „related_to“ verbindet zwei Topics, eines in der Rolle „erstes Thema“ und eines in der Rolle „zweites Thema“, und hat eine Facet „Gewicht“ mit dem Wert 0,6.

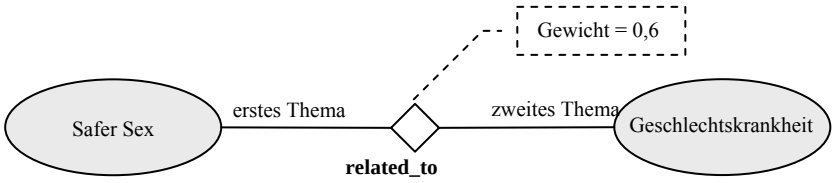


Abbildung 16: Gewichtete related_to-Beziehung in einer Topic Map

Trotz der Verbesserungen im Vergleich zu Semantischen Netzen und der Standardisierung auf XML-Basis sind Topic Maps dem hier behandelten Problem der Modellierung von Themen nicht angemessen: Viele der Konzepte von Topic Maps blieben ungenutzt (Scopes, Occurrences) und es bestünde nach wie vor das bei den Semantischen Netzen erwähnte Problem der beliebig modellierbaren Beziehungen: Entweder der Nutzer würde auf bestimmte vom System vordefinierte Beziehungstypen eingeschränkt (part_of, related_to, …) und dadurch die Ausdrucksstärke der Topic Maps immer weiter beschnitten, oder man erlaubt dem Nutzer zwar neue Beziehungstypen, fordert jedoch eine Modellierung derselben als Association-Type-Topics mit standardisierten Facets, die Informationen zur Behandlung dieser neuen Beziehungstypen durch den Algorithmus enthalten (vgl. Abbildung 17). Letztere Möglichkeit eröffnet interessante Perspektiven, soll hier jedoch nicht weiter verfolgt werden. Ein Algorithmus, der ein beliebig kompliziertes Netzwerk mit n Themen und m Beziehungen, wobei diese Beziehungen von k verschiedenen Beziehungstypen sind, verarbeitet, wird im Rahmen dieser Arbeit als zu komplex angesehen.

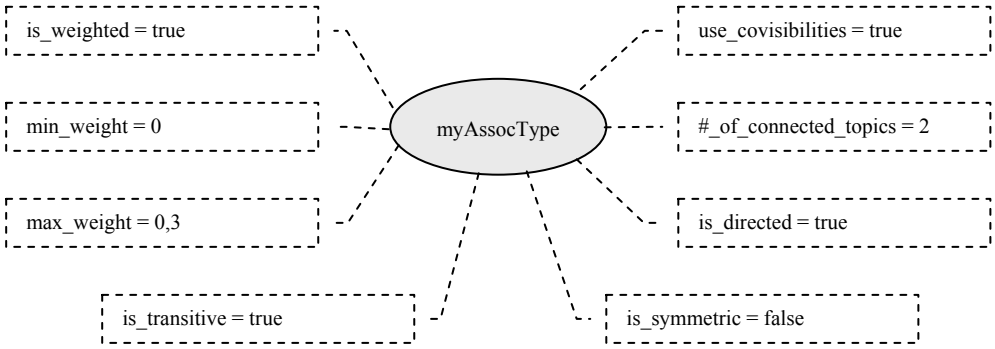


Abbildung 17: Selbst definierter Association Type als Topic mit Facets in einer Topic Map

Topic Maps lassen sich neben der allgemeinen Wissensrepräsentation insbesondere auch zur Navigation und Suche in Internetressourcen sowie zur Formalisierung und dem Austausch von Metadaten verwenden. Auf letzterem liegt auch der Fokus des im Folgenden umrissenen Konzepts, des Resource Description Frameworks.

4.2.3 Resource Description Framework (RDF)

Das Resource Description Framework (RDF, eine Einführung gibt Miller in [Miller98]) ist ein vom W3C standardisiertes ([RDFSpec99]) Modell zur Repräsentation von Metadaten. Einsatz findet es insbesondere im Rahmen des Semantic Web ([BerHenLas01]. Es kann zwischen dem RDF-Model und seiner Serialisierung, der RDF-Syntax (meist in XML, vgl. [XML1.0]), unterschieden werden.

Das für die Modellierung relevante RDF-Model beschreibt Ressourcen an Hand von (Subjekt, Prädikat, Objekt)-Tripeln: Die Subjekte dieser Tripel sind die zu beschreibenden Ressourcen; folgt man der hinter RDF stehenden Grundidee, so sind dies meist (Web-)Dokumente. Schon hier zeigt sich, dass RDF sicher nicht die erste Wahl bei der Modellierung von Themen sein kann: Eine Modellierung von beliebigen Objekten ist zwar auch möglich, widerspricht jedoch der Metapher von RDF. Prädikate sind Eigenschaften (z.B. „Autor“) und Objekte die Werte dieser Eigenschaften (z.B. „Stephen King“), wobei als Objekte wiederum Ressourcen erlaubt sind. Diese Tripel (genannt Statements) ermöglichen die Darstellung binärer Relationen, höherstellige Relationen sind jedoch lediglich über den Umweg der Reification möglich (Aufspaltung der Relation in mehrere Teilrelationen und Einführung von Hilfsobjekten, ähnlich zu Abbildung 15).

Auf RDF soll hier nicht noch weiter eingegangen werden, da es wie gesagt primär zur Annotation von Metadaten zu Webressourcen gedacht ist. Mück/Widhalm formulieren dies im Rahmen eines Vergleichs zwischen Topic Maps und RDF folgendermaßen:

> *„Es handelt sich bei diesen beiden Standards also um zwei an sich unterschiedliche Metamodelle, die einen ähnlichen Zweck verfolgen. Jedoch lässt sich feststellen, dass sich das Modell des RDF eher an der reinen Beschreibung und Charakterisierung von Ressourcen orientiert, während man mit Topic Maps versucht, ein semantisches Netz aufzubauen und Wissensmanagement zu betreiben.“* (vgl. [MüWid02], S. 20)

Zu erwähnen bleibt an dieser Stelle noch

> *„die Schemasprache RDF-Schema, um Definitionen und semantische Einschränkungen (Constraints) für Klassen von RDF-Objekten zu erstellen“* (vgl. [MüWid02], S. 19).

Mit RDF-Schema (ebenfalls eine W3C-Empfehlung, vgl. [RDFSchemaSpec1.0]) lässt sich festlegen, welche Typen und Werte von Eigenschaften für welche Typen von Ressourcen in einem RDF-Model hinreichend oder notwendig sind. Anders gesagt beschreibt RDF reine Fakten, während RDF-Schema mögliche Klassen und Eigenschaften strukturiert. Dies ist sehr ähnlich zu Ontologiemodellierungssprachen wie der Web Ontology Language (OWL), die im Folgenden (anstelle von genauen Ausführungen zu RDF-Schema) vorgestellt wird.

4.2.4 Web Ontology Language (OWL)

Bei der Web Ontology Language handelt es sich um eine durch das W3C spezifizierte Ontologiemodellierungssprache (vgl. [OWLRef04], [OWLGuide04], [HoPatHar03]), die technisch gesehen auf RDF (vgl. 4.2.3) und XML ([XML1.0]) basiert und sich dadurch in die Ontologieschicht des Semantic Web ([BerHenLas01]) eingliedern soll. Historisch gesehen wurde OWL von Beschreibungslogiken wie SHIQ (vgl. [HoSattTob00]) beeinflusst und hat sich aus den Vorgängern OIL (vgl. [Fensel&al01]) sowie DAML&OIL (vgl. [Connolly&al01], [HoPatHar02]) entwickelt. Weder das Semantic Web noch die historische Entwicklung sollen jedoch hier von Belang sein, sondern vielmehr OWL unter dem Gesichtspunkt einer Eignung als Modellierungssprache für Themen beleuchtet werden.

Die Grundelemente von OWL sind Classes, Properties und Individuals. OWL bietet sehr ausdrucksmächtige Möglichkeiten, um Aussagen über Classes und Properties zu treffen. So kann eine Class (Menge von Individuals) auf verschiedene Weise definiert werden (vgl. [OWLRef04], 3.1)[1]:

- lediglich durch ihren Klassenidentifier, wodurch nichts weiter ausgesagt wird, als dass eine Klasse mit diesem Namen besteht
- durch eine Aufzählung aller Individuals, die in diese Class gehören, z.B. „die Klasse, die {Europa, Asien, Nordamerika, Südamerika, Afrika, Australien} enthält“
- durch Eigenschaftsrestriktionen, die beschreiben, welche Eigenschaften hinreichend bzw. notwendig sind, damit ein Individual zu der Klasse gehört, z.B. „die Klasse aller Individuals, die in der Property Wohnort wenigstens ein Mal den Wert ‚Bamberg’ haben“
- durch Bilden der Schnittmenge von Klassen, z.B. „die Klasse aller Schachspieler geschnitten mit der Klasse aller Deutschen“
- durch die Vereinigung von Klassen
- durch Bilden des Komplements einer Klasse, z.B. „die Klasse aller Individuals, die nicht zur Klasse Deutsche gehören“

Auch über Properties können komplexe Aussagen getroffen werden, beispielsweise Transitivität (z.B. is_subregion_of ist üblicherweise transitiv) oder das Festlegen einer inversen Property (z.B. ist_Kind_von ist invers zu ist_Elternteil_von). Ziel ist es, Inferenz zu ermöglichen und dadurch einerseits die Korrektheit der Modellierung zu testen, und andererseits Anfragen zu beantworten, die nicht explizit in den Daten vorhanden sind (z.B. „Ist Person X ein männlicher deutscher Schachspieler mit einem Wohnort aus einem Bundesland, das an Sachsen angrenzt?“).

OWL bietet je nach Anforderungen des jeweiligen Nutzers drei Sprachebenen von verschiedener Komplexität: „OWL Full“ mit allen Sprachkonstrukten, „OWL DL“ mit Einschränkungen, die die Abbildbarkeit auf Beschreibungslogiken (DL = Description Logic) gewährleisten und die Ontologien somit entscheidbar machen, und schließlich als am meisten eingeschränkte Sprachebene „OWL Lite“. Eine Modellierung von Themen würde lange nicht alle Sprachkonstrukte von OWL nutzen. Gerade die sonst so nützliche Inferenz bringt bei der Sichtbar-

[1] Auf Beispiele in OWL-Syntax wird hier der Einfachheit halber verzichtet, sondern auf natürlichsprachige Beschreibung zurückgegriffen.

keitsanalyse kaum Vorteile. Daher liegt es nahe, die am meisten eingeschränkte Variante „OWL Lite“ in Betracht zu ziehen:

> *„OWL Lite supports those users primarily needing a classification hierarchy and simple constraint features. For example, while OWL Lite supports cardinality constraints, it only permits cardinality values of 0 or 1. It should be simpler to provide tool support for OWL Lite than its more expressive relatives, and provide a quick migration path for thesauri and other taxonomies.”* (vgl. [OWLGuide04], 1.1)

OWL Lite wird also als eine Art Zwischenstufe („migration path”) bei der Überführung von Thesauri in Ontologien empfohlen. An dieser Stelle stellt sich die Frage, ob Thesauri dem Problem der Themenmodellierung nicht angemessener sind als ein eigentlich sehr ausdrucksmächtiges OWL, das in der Lite-Version bis zur Unkenntlichkeit beschnitten wurde.

4.2.5 Lexikalisch-semantische Wortnetze und Thesauri

4.2.5.1 Wortnetze in der Computerlinguistik

In 4.2.1 wurde gefordert, die Möglichkeiten der allgemeinen Semantischen Netze auf einige wenige Beziehungen einzuschränken. Ein Ansatz, der genau dies umsetzt, sind (lexikalisch-semantische) Wortnetze (vgl. hierzu [Kunze01]). Im Unterschied zu Semantischen Netzen werden mit Wortnetzen nicht beliebige Objekte, sondern lexikalische Konzepte und ihre Beziehungen zu anderen Konzepten abgebildet, d.h. eine Unterscheidung zwischen Klassen und Instanzen wie in den bisher vorgestellten Ansätzen ist nicht möglich. Dies macht im Bezug auf die Modellierung von Themen auch Sinn, denn die Instanz eines Themas ist oft nicht interpretierbar (z.B. „Instanz des Themas ‚Safer Sex’“).

Das wohl bekannteste Wortnetz ist das von George Miller (et.al.) seit 1985 an der Universität Princeton entwickelte und frei verfügbare WordNet in englischer Sprache (vgl. [Miller&al90], [Miller&al95], [Fellbaum98], http://wordnet.princeton.edu/). In deutscher Sprache entstand an der Universität Tübingen das GermaNet (vgl. http://www.sfs.nphil.uni-tuebingen.de/lsd,

[HampFeld97]), das in das EuroWordNet (vgl. http://www.hum.uva.nl/~ewn/, [Vossen97], [Vossen99]), einem Wortnetz mit 8 europäischen Sprachen, integriert wurde.

Hypernyms:

tennis, lawn tennis -- (a game played with rackets by two or four players who hit a ball back and forth over a net that divides the court)

=> court game -- (an athletic game played on a court)

=> athletic game -- (a game involving athletic activity)

=> sport, athletics -- (an active diversion requiring physical exertion and competition)

=> diversion, recreation --

(an activity that diverts or amuses or stimulates; "scuba diving is provided as a diversion for tourists"; "for recreation he wrote poetry and solved crossw ord puzzles"; "drug abuse is often regarded as a form of recreation")

=> activity -- (any specific activity; "they avoided all recreational activity")

=> act, human action, human activity -- (something that people do or cause to happen)

=> game -- (a contest with rules to determine a winner; "you need four people to play this game")

=> activity -- (any specific activity; "they avoided all recreational activity")

=> act, human action, human activity -- (something that people do or cause to happen)

Hyponyms:

tennis, lawn tennis -- (a game played with rackets by two or four players who hit a ball back and forth over a net that divides the court)

=> professional tennis -- (playing tennis for money)

*=> **singles** -- (tennis played with one person on each side)*

=> doubles -- (tennis played with two players on each side)

=> royal tennis, real tennis, court tennis -- (an ancient form of tennis played in a four-walled court)

Coordinates (Sisters):

tennis, lawn tennis -- (a game played with rackets by two or four players who hit a ball back and forth over a net that divides the court)

-> court game -- (an athletic game played on a court)

=> handball -- (a game played in a walled court or against a single wall by two or four players who strike a rubber ball with their hands)

=> racquetball -- (a game played on a handball court with short-handled rackets)

=> fives -- (a game resembling handball; played on a court with a front wall and two side walls)

=> squash, squash racquets, squash rackets -- (a game played in an enclosed court by two or four players who strike the ball with long-handled rackets)

=> volleyball, volleyball game -- (a game in which two teams hit an inflated ball over a high net using their hands)

=> jai alai, pelota -- (a Basque or Spanish game played in a court with a ball and a wickerwork racket)

=> badminton -- (a game played on a court with light long-handled rackets used to volley a shuttlecock over a net)

=> basketball, basketball game, hoops --

(a game played on a court by two opposing teams of 5 players; points are scored by throwing the basketball through an elevated horizontal hoop)

=> deck tennis -- (game played mainly on board ocean liners; players toss a ring back and form over a net that is stretched across a small court)

=> netball -- (a team game that resembles basketball; a soccer ball is to be thrown so that it passes through a ring on the top of a post)

=> tennis, lawn tennis -- (a game played with rackets by two or four players who hit a ball back and forth over a net that divides the court)

=> pallone -- (an Italian game similar to tennis)

Meronyms:

tennis, lawn tennis -- (a game played with rackets by two or four players who hit a ball back and forth over a net that divides the court)

HAS PART: footfault -- (a fault that occurs when the server in tennis fails to keep both feet behind the baseline)

HAS PART: return -- (a tennis stroke that returns the ball to the other player; "he won the point on a cross-court return")

HAS PART: service break -- (a tennis game won on the opponent's service)

*HAS PART: **advantage** -- (first point scored after deuce)*

HAS PART: set point -- (the final point needed to win a set in tennis)

Abbildung 18: Ausschnitt für „tennis“ aus WordNet (http://wordnet.princeton.edu/)

Konzepte in Wortnetzen entsprechen nicht Wörtern, sondern fassen vielmehr mehrere synonyme Wörter zu einem so genannten Synset zusammen. Dadurch finden Wortnetze beispielsweise Verwendung im Information Retrieval zur Erhöhung der recall-Werte durch Ausnutzen der Synsets, d.h. eine Anfrage nach einem Wort berücksichtigt auch Dokumente, die Synonyme beinhalten. Eine weitere wichtige Einsatzmöglichkeit besteht in der Berechnung der semantischen Nähe zwischen zwei Konzepten über ihre Entfernung zueinander im Wortnetz, was bei der semantischen Analyse von Sätzen in der Computerlinguistik hilfreich ist.
Fragt man beispielsweise nach der semantischen Nähe von „advantage“ und „singles“, so wird man zunächst eine große semantische Entfernung unterstellen oder zumindest überlegen, welche Vorteile nicht in Partnerschaft lebende Personen haben. Doch die Berücksichtigung verschiedener Lesarten führt zu einer gewissen Nähe wie Abbildung 18 (Ausschnitt aus WordNet für „tennis“) zeigt: „advantage“ ist unter anderem nämlich ein Meronym[1] von Tennis, „singles“ ein Hyponym.

Auch für das hier zu lösende Problem der Modellierung von Themen zur Sichtbarkeitsanalyse sind Wortnetze eine in Erwägung zu ziehende Alternative: Wie bereits erwähnt macht es durchaus Sinn, Themen eher als lexikalisch-semantische Konzepte aufzufassen, als sie in das objektorientierte Schema von Klassen und Instanzen pressen zu wollen. Entscheidend für die Eignung ist die Angemessenheit der typischen in Wortnetzen verwendeten Beziehungstypen, die in 4.2.5.3 vorgestellt werden. Vorher jedoch sollen kurz der Begriff Thesaurus und sein Ursprung in der Dokumentationswissenschaft geklärt werden.

4.2.5.2 Thesauri in der Dokumentationswissenschaft

Eine Art von Wortnetzen sind Thesauri: Die beiden Begriffe Thesaurus und Wortnetz sind nicht scharf getrennt und werden oft gleichbedeutend verwendet, wobei der Begriff Thesaurus eher aus der Dokumentationswissenschaft stammt und schon lange vor den computerlinguistischen Wortnetzen verwendet wurde. Thesauri wurden außerdem von DIN und ISO standardisiert (vgl. [ISO2788], [DIN1463]).

[1] Die hier erwähnten Beziehungstypen Meronymie und Hyponymie werden in 4.2.5.3 näher untersucht.

Eine sehr gute Einführung in Thesauri aus dokumentationswissenschaftlicher Sicht gibt [Wersig78]. Wersig vergleicht Thesauri mit anderen Dokumentationssprachen und ordnet sie eher als natürlich-sprachlich basiert ein, wodurch sie sich besonders für den systemspezifischen Einsatz eignen (vgl. Abbildung 19) – ein wesentlicher Unterschied zu den bisher vorgestellten Modellierungsansätzen wie Semantischen Netzen, Topic Maps oder OWL, die eher auf eine allgemeine Anwendbarkeit abzielen.
Im Rest dieser Arbeit soll „Thesaurus“ gleichbedeutend mit „Wortnetz“ verwendet werden.

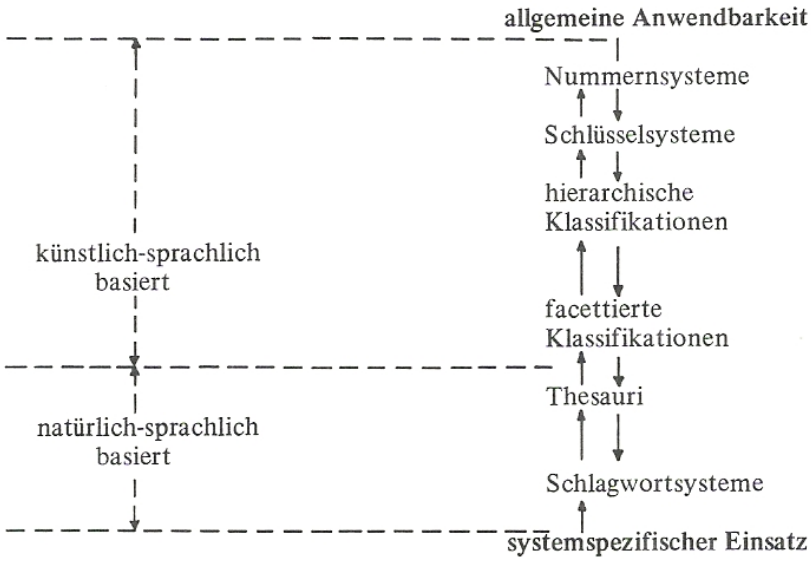


Abbildung 19: Kontinuum der Verwendung von Dokumentationssprachen (vgl. [Wersig78], Seite 30, S8)

4.2.5.3 Beziehungstypen in Wortnetzen

Dieser Abschnitt beschreibt die für Thesauri typischen Beziehungstypen und orientiert sich vornehmlich an den Ausführungen von [Kunze01] zu GermaNet und an den in [Wersig78], Punkt 3.5 strukturierten Beziehungstypen. Weitere Ausführungen finden sich in der Literatur zur lexikalischen Semantik (vgl. insb. [Cruse95], [JurMartin00] Kap. 16). Für jeden Beziehungstyp soll kurz seine Eignung für die Sichtbarkeitsanalyse diskutiert werden.

(1) *Synonymie* wurde bereits in 4.2.5 erwähnt und ist Bedeutungsgleichheit zwischen Wörtern (nicht zwischen Konzepten), z.B. „Geldbeutel“, „Portemonnaie“. Wie bereits beschrieben ist Synonymie in GermaNet, WordNet und anderen Wortnetzen bereits durch die Model-

lierung der Konzepte als Synsets realisiert, so dass im Wortnetz keine Kante für Synonymie vorgesehen ist.

Selbstverständlich können Themen synonym zueinander sein, beispielsweise „Präservativ“ und „Kondom“. Da Themen in dieser Arbeit nur durch ihren Namen beschrieben werden (vgl. 2.1) und keine weiteren Eigenschaften wie die Menge aller Synonyme zuordenbar sind, lassen sich Synsets nicht darstellen, so dass eine Kante „synonym“ vorgesehen werden sollte. Abbildung 20 zeigt die im Folgenden hierfür verwendete Darstellung:

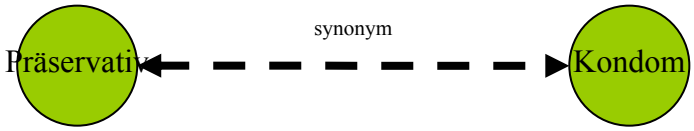


Abbildung 20: Synonymie im Wortnetz

(2) *Antonymie* ist die zur Synonymie konverse Relation, die ein Gegenteil ausdrückt.

Auch Antonymie kann zwischen Themen bestehen, beispielsweise „Safer Sex“ und „Unsafe Sex“, und kann in bestimmten Fällen auch einen Einfluss auf den Sichtbarkeitsausgleich haben. So wird auf einer Seite, die „Safer Sex“ behandelt, zu einem gewissen Grad auch „Unsafe Sex“ thematisiert (ohne dass notwendigerweise der Begriff „Unsafe Sex“ vorkommen muss), denn gerade diese Sexualpraktik soll durch Thematisierung von „Safer Sex“ verhindert werden.

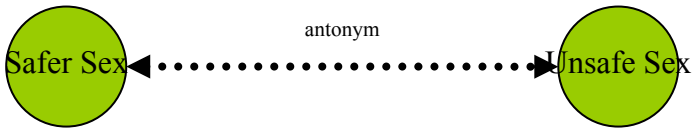


Abbildung 21: Antonymie im Wortnetz

(3) *Hyponymie* entspricht der ebenfalls bereits erwähnten is_a-Beziehung, d.h. einer Bildung von Ober- und Unterbegriffen, wie sie zur Klassifikation[1] verwendet wird. *Hyperonymie* ist in diesem Fall die konverse Relation, muss jedoch nicht eigens modelliert werden, da sie durch die Hyponymie bereits impliziert wird.

Hyponymie kann zwischen Themen bestehen, beispielsweise „Geschlechtskrankheit“, „Syphilis“ und „Aids“.

[1] Klassifikation ist ein besonderer Typ von Wortnetz, der lediglich Hyponymien/Hyperonymien verwendet und somit eine Baumstruktur schafft, vgl. z.B. [Wersig78] Punkt 1.3

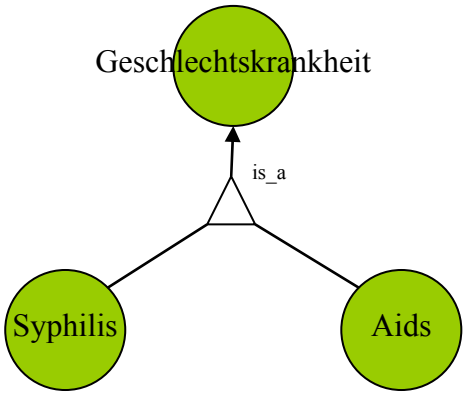


Abbildung 22: Hyponymie im Wortnetz

(4) *Meronymie* wird auch als Partonomie-, Teil/Ganzes- oder part_of-Beziehung bezeichnet. Die konverse Relation ist die *Holonymie*. Meronymie und Hyponymie sind transitiv.
Meronymie kann ebenfalls zwischen Themen auftreten, beispielsweise ist „Finanzminister“ ein Teil von „Kabinett“. Es ist Ansichtssache, ob die Sichtbarkeit von „Kabinett“ durch eine Diskussion über „Finanzminister“ erhöht wird. Grundsätzlich sollte dafür eine Möglichkeit geschaffen werden, so dass auch die Meronymie zur Modellierung von Themen verwendet werden soll.

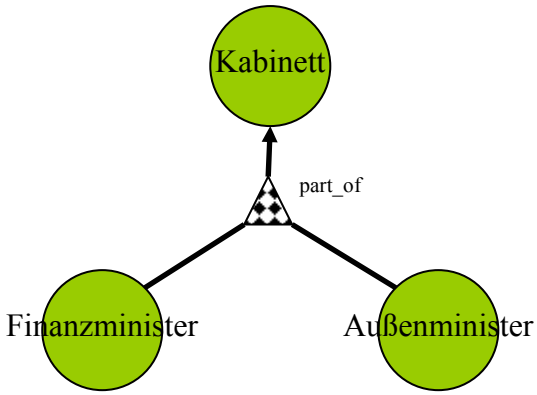


Abbildung 23: Meronymie im Wortnetz

(5) Andere gerichtete Begriffsbeziehungen (vgl. [Wersig78] Kap. 3.5.2.6.) wären beispielsweise genetische Beziehung, Vorgänger-Nachfolger-Beziehung, Urheber-/Verursacherbeziehung, Materialbeziehung, Kausalbeziehung (Kausation in GermaNet), zeitlicher Zusammenhang oder Implikation (GermaNet).
Da diese zwischen Themen jedoch selten sind und die Komplexität unnötig erhöhen, soll wie bei vielen Thesauri verfahren werden und anstelle dieser Beziehungen eine *Assoziation* verwendet werden:

„Obwohl es eine Reihe weiterer gerichteter Begriffsbeziehungen gibt, sind diese in Thesauri meist nicht getrennt ausgewiesen, sondern werden im Einzelfall bei Bedarf als assoziative Beziehung dargestellt.“ (vgl. [Wersig78], Seite 134)

Eine Assoziation zwischen Themen sagt also nur aus, dass diese Themen in einer (nicht weiter definierten) Beziehung zueinander stehen. Das kann eine Kausalbeziehung („Mord“ verursacht „Tod“) oder ein beliebiger anderer thematischer Zusammenhang sein.

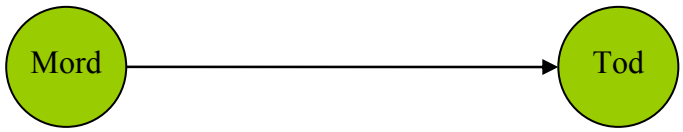


Abbildung 24: Assoziation im Wortnetz

Verwendet man nun einen Thesaurus mit Synonymie, Antonymie, Hyponymie, Meronymie und Assoziation zur Modellierung von Themen, so ergeben sich in Bezug auf die in 3.1 formulierten Anforderungen folgende Ergebnisse:

- Es ist möglich, gerichtete Beziehungen zu modellieren (außer bei den symmetrischen Beziehungen Synonymie und Antonymie).
- Die Modellierung ist für den Benutzer äußerst intuitiv handhabbar.
- Der Modellierungsoutput ist weniger komplex als der einer Topic Map oder von OWL, doch durch fünf verschiedene Beziehungstypen immer noch komplexer als das Themennetzwerk aus 4.1
- Es ist nicht möglich, die Stärke einer Beziehung auszudrücken.

Das Beispiel aus Abbildung 13 lässt sich als Thesaurus wie folgt darstellen (vgl. Abbildung 25):

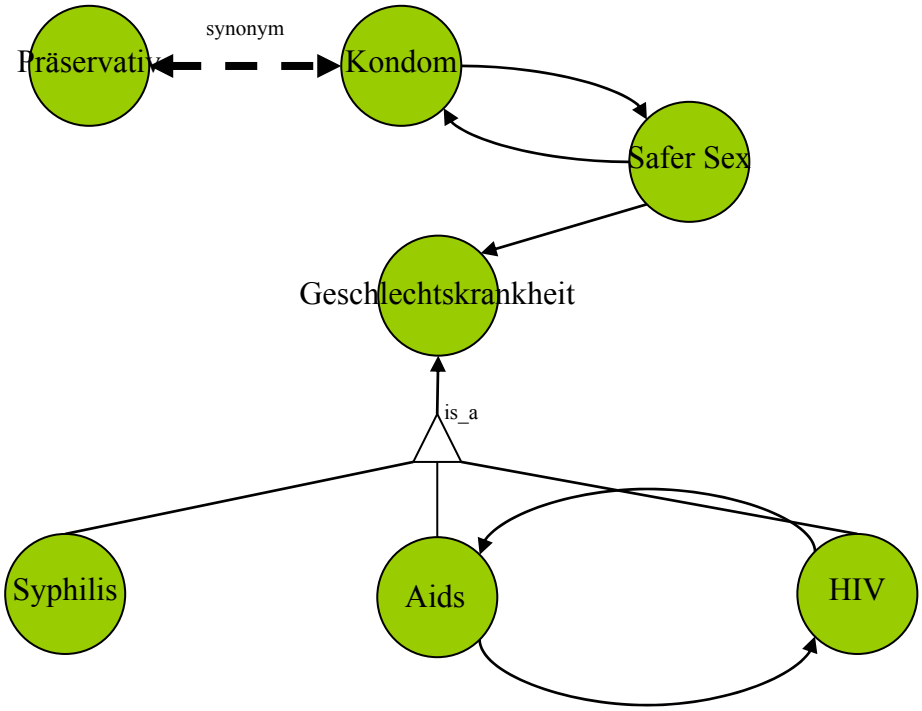


Abbildung 25: Thesaurus zum Themengebiet „Safer Sex“

4.3 Auswahl eines Modellierungsansatzes

4.3.1 Vergleich der Modellierungsansätze

Tabelle 3 fasst die bisherigen Überlegungen zur Modellierung der semantischen Beziehungen zwischen Themen an Hand der in 3.1 gelisteten Kriterien für eine geeignete Modellierung zusammen: Keiner der Ansätze überzeugt vollständig. Der Einsatz eines Ansatzes, der dem Problem nicht angemessen ist, scheidet auf jeden Fall aus, so dass in einer ersten Vorauswahl die drei verbleiben, die dem Problem am ehesten angemessen sind: Themennetzwerk, Semantisches Netz und Thesaurus. Das Semantische Netz kommt auf Grund seiner beliebigen Beziehungen nicht in Frage und außerdem behebt der Thesaurus (als eine Form von Semantischem Netzes) gerade dieses Problem durch die Beschränkung auf fünf Beziehungen. Es verbleiben das Themennetzwerk, das keine komplexen Beziehungen unterstützt, und der Thesaurus, der keine Gewichtungen zulässt – beide verletzen unbedingt notwendige Anforderungen und sind so nicht einsetzbar. Eine Idee wäre die Kreation eines kombinierten Ansatzes, das heißt ein modifizierter Thesaurus, der Kantengewichte zulässt. Andererseits bestünde auch die Möglichkeit, den Nutzer einen Thesaurus modellieren zu lassen und in einem zwei-

ten Schritt daraus ein Themennetzwerk abzuleiten. Ein großer Vorteil wäre hierbei auch die Reduktion der Beziehungstypen von fünf beim Thesaurus auf nur eine beim Themennetzwerk, was den folgenden Algorithmus immens vereinfacht. Aus diesem Grund soll diese zweite Möglichkeit vorgezogen werden.

		Themen-netzwerk	**Semantisches Netz**	**Topic Map**	**RDF**	**OWL**	**Thesau-rus**
gerichtete Beziehungen		ja	ja	ja	ja	ja	ja
Stärke der Beziehungen		ja	nur durch Reification	ja	nur durch Reification	ja	nein
Intuitive Model-lierung	**komplexe Bezieh.**	nein	ja	ja	ja	ja	ja
	angemesse-ner Ansatz*	ja	ja	fast**	nein	nein	ja
Komplexität d. Outputs (Anzahl an möglichen Beziehungstypen)		1	beliebig	beliebig***	beliebig****	beliebig	5

* Passende Metapher? Sind die Konzepte des Ansatzes dem Problem angemessen oder bleiben wesentliche Möglichkeiten ungenutzt?

** Occurences und Scopes bleiben ungenutzt.

*** aber Beziehungen können in Topic Map definiert werden

**** durch RDF-Schema einschränkbar

Tabelle 3: Übersicht: Verschiedene Modellierungsansätze und ihre Eignung zur Modellierung von Themen

4.3.2 Ableitung eines Themennetzwerks aus einem Thesaurus

Die Ableitung eines Themennetzwerks aus einem gegebenen Thesaurus bedeutet die Abbildung der fünf verschiedenen Beziehungstypen Synonymie, Antonymie, Hyponymie, Meronymie und Assoziation auf nur einen Beziehungtyp, die einfache Assoziation. Es ist offensichtlich, dass dies in jedem Fall nur teil-automatisiert durchgeführt werden kann, da die Unterstützung des Nutzers zumindest bei der Festlegung von Gewichten notwendig ist.

4.3.2.1 Vorgehensweise

Folgendes schrittweises Vorgehen wird vorgeschlagen:

1) Zusammenfassen der Synonyme (vollautomatisiert)

   Alle Themen, die im Thesaurus untereinander mit Synonymie-Kanten verbunden sind, werden zu einem Thema zusammengefasst. Wie in 4.2.5.3 bereits beschrieben, sind Synsets in dieser Arbeit eigentlich nicht darstellbar, da jedes Thema nur durch seinen Namen repräsentiert wird (vgl. 2.1). Daher soll im Folgenden in Anlehnung an die übliche Syntax von Anfragen an Suchmaschinen ein Thema, das mehrere synonyme Begriffe hat, als Verknüpfung dieser Begriffe mit „OR“ repräsentiert werden.

   Die Beziehungen, die zwischen je einem der zusammengefassten Themen zu anderen Themen außerhalb des Synsets werden übernommen (vgl. Abbildung 26)

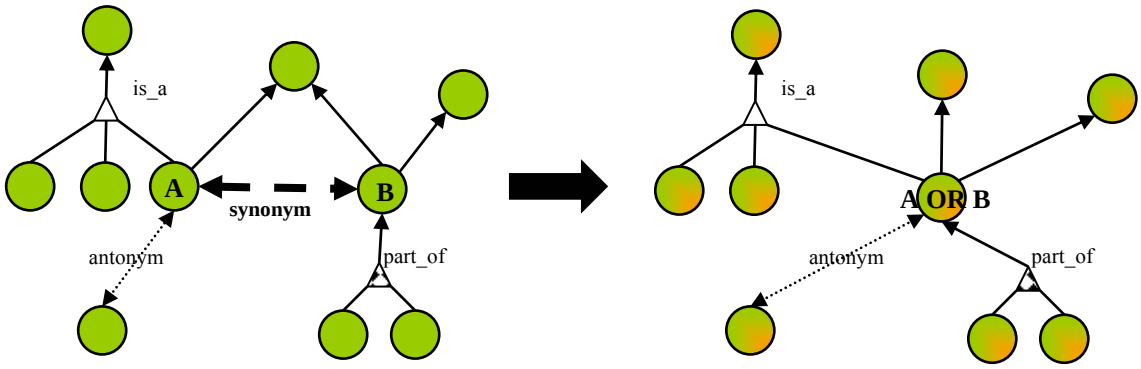


Abbildung 26: Ableitung eines Themennetzwerks aus einem Thesaurus: Eliminierung von Synonymie

2) Umwandeln von Antonymie in Assoziation (teilautomatisiert)

   Themen, die antonym zueinander sind, können im Themennetzwerk je eine Assoziationsbeziehung in jede Richtung zwischen sich haben (vgl. Abbildung 27). Da dies für jede Antonymie unterschiedlich ist, muss an dieser Stelle der Nutzer eingreifen und zunächst entscheiden, ob zwei, eine oder keine Assoziation im Themennetzwerk vorhanden sein sollen, und weiterhin, wie stark diese Beziehung(en) ist/sind. Üblicherweise dürften diese Beziehungen eher mit kleinen Gewichten versehen sein; beispielsweise erhöht eine Diskussion über „Freiheit“ nur in bestimmten Fällen (je nach Kontext) die Sichtbarkeit des Antonyms „Unfreiheit“.

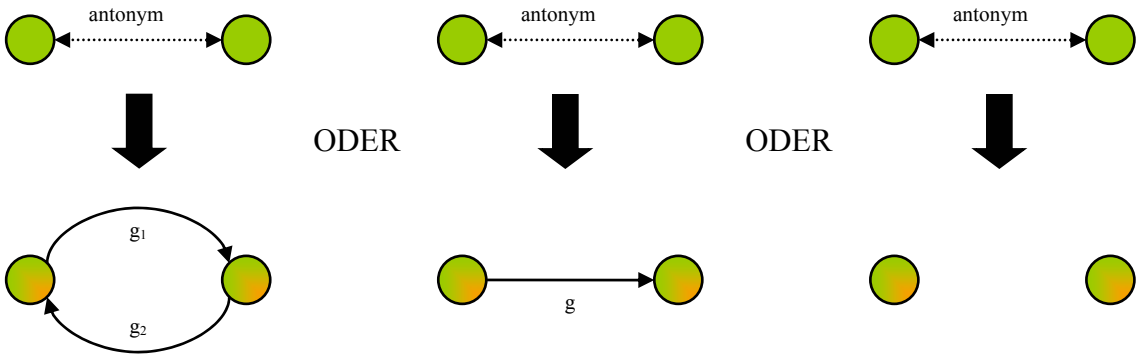


Abbildung 27: Ableitung eines Themennetzwerks aus einem Thesaurus: Umwandeln von Antonymie

3) Umwandeln von Hyponymie in Assoziation (teilautomatisiert)

Hyponymien werden zu Assoziationen, wobei von jedem Unterthema zu seinem Oberthema eine Assoziation erstellt wird (vgl. Abbildung 28). Das Gewicht dieser Assoziation ist im Normalfall sehr hoch. Assoziationen in die Gegenrichtung sind nicht vorgesehen, können jedoch dadurch realisiert werden, dass im Thesaurus sowohl eine is_a-Beziehung, als auch eine Assoziation modelliert wurde.

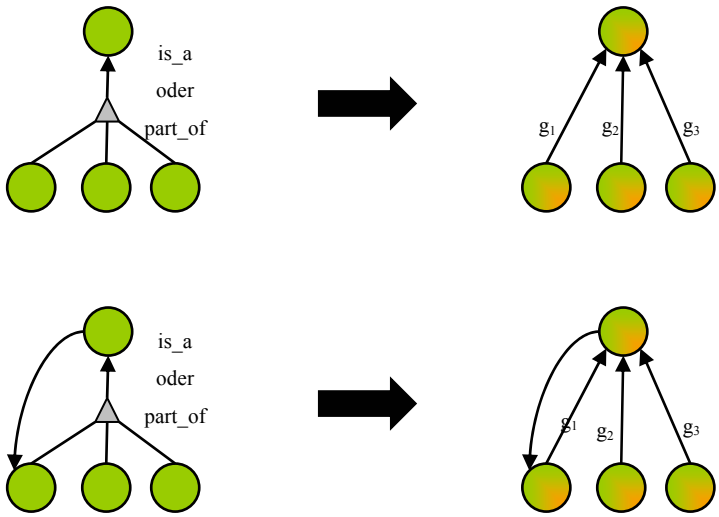


Abbildung 28: Ableitung eines Themennetzwerks aus einem Thesaurus: Umwandeln von Hyponymie oder Meronymie

4) Umwandeln von Meronymie in Assoziation (teilautomatisiert)

Es wird entsprechend wie bei der Hyponymie vorgegangen (vgl. Abbildung 28), doch die Gewichte sind im Allgemeinen niedriger als bei der Hyponymie.

5) Gewichten der bereits bestehenden Assoziationen (nicht automatisiert)
   Die Assoziationsbeziehungen aus dem Thesaurus werden mit Gewichten versehen. Eine Idee, den Nutzer bei dieser Gewichtung zu unterstützen, soll in 4.3.2.2 erwähnt werden.

Der hohe Anteil an nicht- oder teilautomatisierten Schritten wirft zu Recht die Frage auf, ob eine direkte Modellierung des Themennetzwerks nicht weniger Aufwand als das hier vorgeschlagene Vorgehen wäre. Doch wollte man ein Themennetzwerk mit nur 10 Themen direkt modellieren, so müsste man für 90 (= 10 · 9) mögliche Kombinationen überlegen, ob eine irgendwie geartete Beziehung zwischen zwei Themen besteht, während bei der Ableitung des Thesaurus schon eine Vorauswahl der möglichen Beziehungen getroffen wurde. Außerdem entspricht das Denken in den komplexen Beziehungen bei der Erstellung des Thesaurus mehr der menschlichen Denkweise („Ist X ein Unterbegriff von Y?“ ist schneller zu beantworten als „Steht X in irgendeiner Beziehung zu Y?“).
Das mit oben stehenden Regeln abgeleitete Themennetzwerk zum Thesaurus aus Abbildung 25 wird hier in keiner Abbildung veranschaulicht. Es unterscheidet sich von der ersten Modellierung eines Themennetzwerkes in Abbildung 13 nur geringfügig durch die Zusammenfassung der synonymen Themen „Präservativ“ und „Kondom“ zu einem Knoten „Präservativ OR Kondom“.

4.3.2.2 Unterstützung der Gewichtsfestlegung durch Covisibilitäten

Automatisiert festzustellen, ob eine semantische Beziehung zwischen zwei Themen besteht und wie stark diese Beziehung ist, ist nicht möglich. Doch ein Anhaltspunkt lässt sich eventuell an den Covisibilitäten der Themen im Internet suchen: Zwei Themen, die sehr stark miteinander semantisch verwandt sind, werden mit großer Wahrscheinlichkeit eine große Anzahl an Webseiten aufweisen, auf denen sie beide vorhanden sind.
Verdeutlicht werden soll das an der Frage, welche der Themen „HIV“, „Aids“ und „Grippe“ in einem semantischen Zusammenhang stehen. Anhand der in 2.3.2.1 definierten Formeln für Co-Sichtbarkeitsmaße wird die Co-Sichtbarkeit (mit Trefferzahlen der Suchmaschine Google™) für je zwei der drei Themen „HIV“, „Aids“ und „Grippe“ berechnet (vgl. Tabelle 4). Die drei ersten Zeilen wurden aus Tabelle 1 übernommen.

covisibility(HIV, Aids)	0,183983
$covisibility_1$(HIV, Aids)	0,397816
$covisibility_2$(HIV, Aids)	0,255000
covisibility(HIV, Grippe)	0,002157
$covisibility_1$(HIV, Grippe)	0,002620
$covisibility_2$(HIV, Grippe)	0,044342
covisibility(Aids, Grippe)	0,002310
$covisibility_1$(Aids, Grippe)	0,002380
$covisibility_2$(Aids, Grippe)	0,072783

Tabelle 4: Co-Visibilitäten der Themen „Aids“, „HIV“ und „Grippe“

Alle drei Werte sind für (HIV, Aids) bei weitem höher als bei (HIV, Grippe) oder (Aids, Grippe). Welcher der drei Covisibilitätswerte sinnvollerweise eingesetzt werden sollte, soll gar nicht diskutiert werden, sondern nur darauf hingewiesen, dass die Covisibilitätswerte in vielen Fällen einen Hinweis auf mögliche semantische Beziehungen zwischen Themen und deren Stärke geben können.

5. ALGORITHMISCHER AUSGLEICH DER SICHTBARKEITEN VON THEMEN

Output des in 4. beschriebenen Modellierungsvorganges ist ein Themennetzwerk (vgl. 4.1), d.h. ein gerichteter Graph (für Beispiele vgl. Abbildung 12 und Abbildung 13):

$$\textbf{topicrelationships}: \text{TOP} \times \text{TOP} \rightarrow [0..1]$$

mit: topicrelationships(a, a) = 0

Gewichtet sind die Knoten des Themennetzwerks mit den initialen Sichtbarkeiten, die im Folgenden mit visibility_0 (Index 0 für den Zeitpunkt) bezeichnet werden:

$$\textbf{visibility}_\mathbf{0}: \text{TOP} \rightarrow \text{IR}^+_0$$

Aufgabe des algorithmischen Sichtbarkeitsausgleichs ist es nun, daraus derart abgeleitete Sichtbarkeiten **visibility'**: TOP → IR^+_0 zu berechnen, dass die in 3.2 formulierten Anforderungen beachtet werden:

(1) Keine Abnahme von Sichtbarkeiten
(2) Unterstützung transitiver Beziehungen
(3) Stabilität gegenüber kleinen Veränderungen der Kantengewichte
(4) Stabilität gegenüber kleinen Veränderungen der Anfangswerte
(5) Berücksichtigung von Covisibilitäten
(6) Ergebnisse, die im Rahmen des intuitiv Sinnvollen liegen (z.B. nicht unangemessen hoch)

In 5.2 werden hierzu schrittweise vier verschiedene Algorithmen erarbeitet, deren Entstehung zu Teilen von zwei aus verschiedenen Richtungen stammenden und in 5.1 vorzustellenden Ansätzen – dem PageRank-Algorithmus und der Spreading Activation – inspiriert wurde.

5.1 Ansätze aus der relevanten Fachliteratur

5.1.1 PageRank-Algorithmus

Der PageRank-Algorithmus (vgl. [Page&al98], [BrinPage98]) ist einer der Hauptbestandteile der Suchmaschine Google™ (http://www.google.com) und liefert eine Antwort auf die Frage, wie die Linkstruktur in einer Hypertext-Umgebung wie dem Internet genutzt werden kann, um die für eine Suchanfrage des Nutzers als relevant erachteten Dokumente in einer geeigneten Reihenfolge zu präsentieren. Diese geeignete Reihenfolge richtet sich nach der Wichtigkeit der Seiten; PageRank versucht, jeder Webseite ein Wichtigkeitsmaß auf Grund der Linkstruktur zuzuweisen (vgl. Abbildung 29).

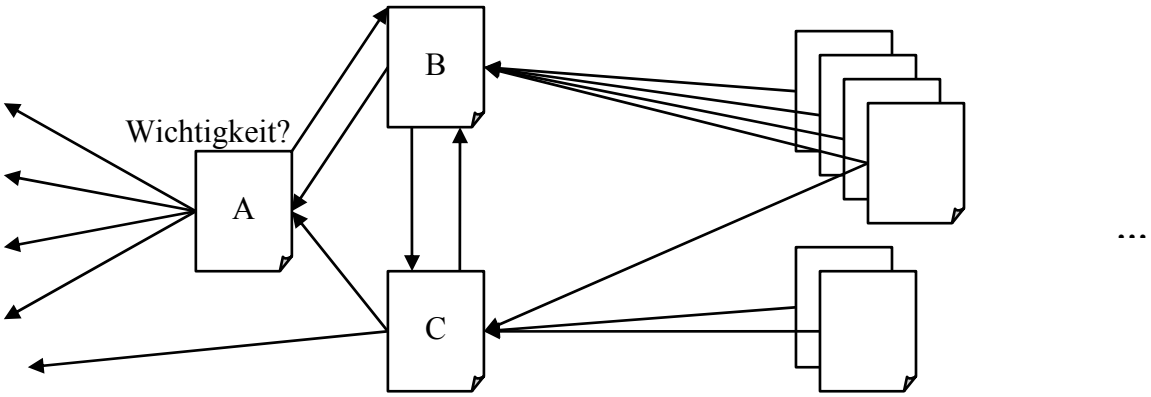


Abbildung 29: Linkstruktur zwischen Webdokumenten

Beispielsweise sollen bei einer Anfrage nach „Java“ die den Begriff „Java“ enthaltenden Webseiten weder in einer zufälligen Reihenfolge gelistet werden, noch soll diejenige Webseite präferiert werden, die den Begriff „Java“ am häufigsten enthält[1], sondern möglichst die Java-Hauptseite von Sun® als die für die Programmiersprache Java relevanteste Seite[2].

Wichtig ist eine Seite nach PageRank dann, wenn andere wichtige Seiten auf sie verweisen:

---

[1] Ein Ranking an Hand der Häufigkeit eines Suchbegriffes in einem Dokument würde Versuchen, sich durch „unlauteres Vorgehen“ im Ranking der Suchmaschinen nach oben zu drängen und damit Aufmerksamkeit auf sich zu ziehen (vgl. 1.3.1, Aufmerksamkeitsökonomie), Tür und Tor öffnen.

[2] Oder wahlweise eine sehr wichtige Seite über die indonesische Insel Java oder die Kaffeesorte Java, doch dieses semantische Problem der Homonymie fällt in den hier nicht weiter diskutierten Bereich des Semantic Web (vgl. [BerHenLas01]).

„a page has high rank if the sum of the ranks of its backlinks is high" ([Page&al98], Seite 3)

Dadurch ist zur Ermittlung des Wichtigkeitsmaßes von Seite X nicht nur die Menge aller auf X verweisenden Seiten P_1 relevant, sondern auch die Menge P_2 aller Seiten, die auf ein Element aus P_1 verweisen, was sich beliebig bis zur transitiven Hülle P^* fortsetzen lässt, so dass letztlich der gesamte Teilgraph aller Knoten, von denen ein Pfad zu X existiert, mit seiner (im Allgemeinen auch zyklischen) Linkstruktur zu berücksichtigen ist.

Der PageRank einer Seite wird folgendermaßen berechnet:

„We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. […] Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one." (vgl. [BrinPage98], Seite 4)

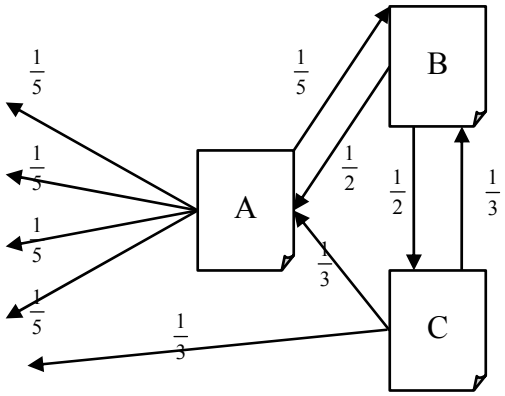


Abbildung 30: Linkstruktur zwischen Webseiten mit Link-Gewichtungen nach PageRank

Anschaulich beschrieben bedeutet dies, dass jede Webseite einen Bruchteil ihrer Wichtigkeit an jede Seite weitergibt, auf die sie verweist (genauer: an jede verlinkte Seite den gleichen Anteil). Abbildung 30 verdeutlicht diesen Zusammenhang an Hand eines Teilausschnitts aus Abbildung 29. Die Gleichgewichtung aller ausgehenden Links beruht auf dem Random Surfer Model (vgl. [Page&al98] Kap. 2.5), das davon ausgeht, dass ein Surfer zufällig einen Link auswählt, was zu einer Gleichwahrscheinlichkeit der ausgehenden Links führt.

Ohne weiter auf Details der PageRank-Formel einzugehen, resultiert daraus ein lineares Gleichungssystem der folgenden Form (d = 0,85):

$$PR(A) = 0{,}15 + 0{,}85 \cdot (\tfrac{1}{2} PR(B) + \tfrac{1}{3} PR(C))$$

$$PR(B) = 0{,}15 + 0{,}85 \cdot (\tfrac{1}{5} PR(A) + \tfrac{1}{3} PR(C))$$

$$PR(C) = 0{,}15 + 0{,}85 \cdot (\tfrac{1}{2} PR(B))$$

Dieses System ist sehr einfach lösbar, doch in realen PageRank-Situationen liegt die Anzahl der Seiten nicht bei drei wie hier, sondern eher bei mehreren hundert Millionen ([Broder&al00], Seite 1: „200 million pages and 1.5 billion links“ im Jahr 2000), so dass eine Berechnung durch einen iterativen Algorithmus der Lösung des Gleichungssystems durch direkte Verfahren der Linearen Algebra vorzuziehen ist. Dieser iterative Algorithmus weist jeder der Seiten einen beliebigen Startwert (initialer PageRank) zu und berechnet anschließend sooft oben stehende Gleichungen und normiert das Ergebnis auf 1, bis die Werte konvergieren. Der Algorithmus liefert also unabhängig von den Startwerten das gleiche Ergebnis, im Beispiel vgl. Tabelle 5.

Seite	PageRank (20 Iterationen)
A	0,386752474
B	0,315160254
C	0,298087272

Tabelle 5: Ergebnis des PageRank-Algorithmus für Abbildung 30

Der Ausgangspunkt von PageRank erinnert sehr an das Problem des Sichtbarkeitsausgleichs: Beide Probleme arbeiten auf gerichteten Graphen, in beiden Fällen geben Knoten einen Teil ihres (Sichtbarkeits-/PageRank-)Wertes weiter, auch transitive und zyklische Beziehungen werden jeweils berücksichtigt und eine Konvergenz der Werte wie bei PageRank ist auch beim Sichtbarkeitsausgleich wünschenswert. Es wäre auch ohne weitere Probleme möglich, PageRank für Graphen mit gewichteten Kanten zu verallgemeinern und dann auf ein Themennetzwerk anzuwenden.

Dennoch kommt PageRank für den Sichtbarkeitsausgleich nicht in Frage, da die Ergebnisse unabhängig von den Initialwerten sind, was bei der Sichtbarkeitsanalyse absolut ausgeschlos-

sen ist. Die initialen Sichtbarkeiten sind schließlich keine beliebigen Startwerte wie bei PageRank, sondern eine entscheidende Größe.

5.1.2 Spreading Activation

Spreading Activation wurde in der Psychologie schon in den 60er Jahren von Quillian (vgl. [Quillian68]) als ein Modell für die Aktivierung des menschlichen Gedächtnisses vorgeschlagen und später vielfach eingesetzt und abgeändert (vgl. z.B. [CollLoft75]).

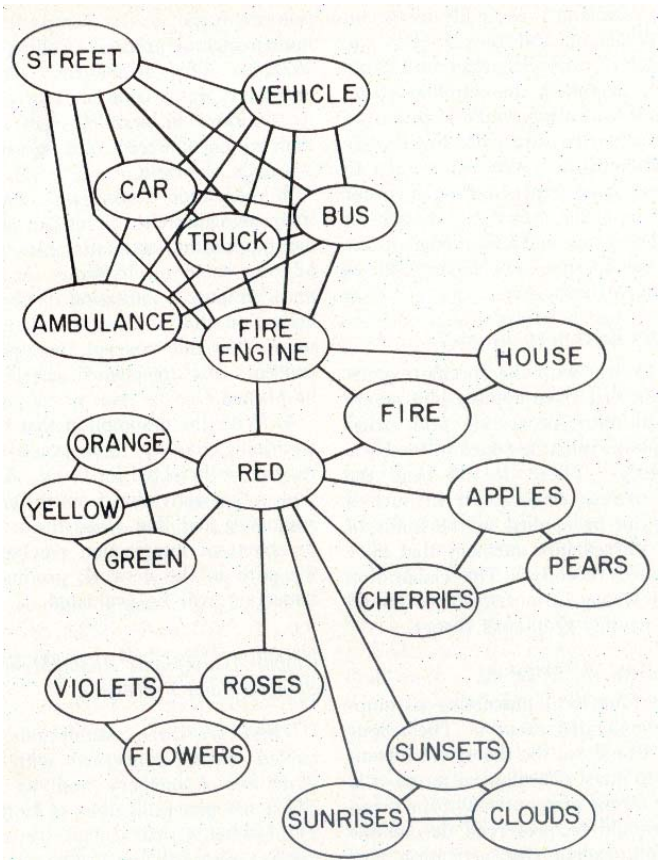


Abbildung 31: Semantisches Netz zur Repräsentation eines Ausschnitts des menschlichen Gedächtnisses in Nachbarschaft zum Konzept „red“ (vgl. [CollLoft75], Figure 1)

Quillian interpretiert das menschliche Gedächtnis als ein Semantisches Netz (vgl. 4.2.1) von Konzepten. Entscheidend ist hier, dass die Beziehungen zwischen Konzepten nicht etwa auf Grund einer lexikalischen Semantik und wie im Thesaurus normativ und möglichst objektiv festgelegt wurden, um die Konzepte zu strukturieren (vgl. 4.2.5), sondern dass eine Beziehung die Verknüpfung zweier Konzepte im individuellen Gedächtnis einer Person repräsen-

tiert. So kann beispielsweise bei Individuum A auf Grund seiner Tätigkeit als Blumenhändler das Konzept „red“ sehr eng mit dem Konzept „roses“ verknüpft sein, während bei Individuum B als Feuerwehrmann „red“ und „fire engine“ näher zueinander sind. Abbildung 31 zeigt ein solches Semantisches Netz aus [CollLoft75], wobei kurze Kanten zwischen Knoten für eine enge Beziehung der verknüpften Knoten stehen, lange Kanten für eine schwache Beziehung (die Kanten sind folglich durch die gewählte Visualisierung gewichtet).

Spreading Activation verfolgt nun das Ziel, die sukzessive Aktivierung verschiedener Konzepte im menschlichen Gedächtnis bei Perzeption eines Konzepts nachzuvollziehen: Wieso denkt Person A bei Erwähnung der Farbe Rot sofort an Rosen und Person B in erster Linie an Feuerwehrautos? Die Metapher des Ansatzes ist ein Energiestoß an einem bestimmten Knoten des Netzes (in diesem Fall der Knoten „red“). Diese Energie verbreitet sich zunächst in einem ersten Schritt zu den direkt benachbarten Knoten, wobei die Anfangsenergie proportional zu den Kantengewichten auf die Nachbarknoten aufgeteilt wird. Im zweiten Schritt verteilt sich diese neu gewonnene Energie auf die gleiche Art und Weise von jedem der Nachbarknoten zu deren Nachbarknoten – hierbei können Knoten auch mehrfach und von verschiedenen Nachbarn Energie erhalten. Auf diese Weise verteilt sich die Energie im gesamten Netz, bis die in einem Schritt von einem Knoten gewonnene Energie unter einem bestimmten Wert liegt. Als Algorithmus lässt sich das folgendermaßen rekursiv formulieren:

„*procedure energize* ($e \in R^{+}_{0}$, $s \in V$) {

energy(s) $\leftarrow$ *energy(s)* + *e;*

$e' \leftarrow e / \sum_{(s,n)\in E} W(s,n)$ *;*

if e > T then

$\forall(s, n) \in E$: *energize* ($e' \cdot W(s, n), n$);

end if

}“ (vgl. [LauZieg04], Seite 7)

V ist hierbei die Menge aller Knoten, E die Menge aller Kanten, W(s,n) das Gewicht der Kante von s nach n und T die Grenze, unter der eine Energiemenge sich nicht weiterverbreitet.

Nach diesem Modell des Gedächtnisses wird ein Konzept, das nach Terminierung des Algorithmus' einen hohen Energiewert hat, von diesem Individuum als sehr nahe zu dem Ursprungskonzept (dem der Energiestoß versetzt wurde) angesehen. Eine weitere Möglichkeit ist die Energiezuführung an zwei Konzeptknoten nacheinander mit späterer Evaluation ob erstens sich die beiden Energiestöße im Laufe der Ausbreitung überschnitten haben und zweitens, auf welchem Pfad dies geschehen ist. So ließe sich beispielsweise erklären, dass ein Individuum „cherries" und „ambulance" mental über den Pfad „cherries" → „red" → „fire engine" → „ambulance" miteinander verbindet.

Auch auf Gebiete außerhalb der Psychologie wurde die Spreading Activation übertragen, von denen hier zwei beispielhaft genannt werden sollen:
So bringt es Preece (vgl. [Preece81]) im Information Retrieval zum Einsatz. Ebenfalls im Information Retrieval schlagen Ceglowski&al (vgl. [CegCobCua03]) Spreading Activation als Alternative zum Latent Semantic Indexing vor. Sie wenden den Ansatz auf Contextual Network Graphs an – auf Graphen, in denen Dokumente mit in ihnen vorkommenden Begriffen verbunden werden – und ermöglichen damit bei Suchanfragen die Berücksichtigung ähnlicher Begriffe sowie die Suche nach Dokumenten, die ähnlich zu einem gegebenen Dokument sind.
Lausen und Ziegler (vgl. [LauZieg04]) schließlich basieren ihre Appleseed Trust Metric zur Berechnung von Vertrauenswerten in Trust-Netzwerken auf Spreading Activation.

Im folgenden Punkt 5.2 soll neben drei anderen Algorithmen auch ein auf Spreading Activation basierender Algorithmus zum Sichtbarkeitsausgleich vorgeschlagen werden.

5.2 Mögliche Algorithmen

5.2.1 SimplePropagation

Die erste und ziemlich triviale Idee zur Durchführung des Sichtbarkeitsausgleichs soll SimplePropagation genannt werden. Dieser Algorithmus eignet sich – wie offensichtliche Überlegungen zeigen werden – nur schlecht für die Sichtbarkeitsanalyse und dient lediglich als Ausgangspunkt für die anschließenden Algorithmen.

Beispiel sei das oben in Abbildung 12 dargestellte Themennetzwerk mit zwei Knoten. Die Semantik der Kante zwischen den beiden Knoten wurde in 3.1 folgendermaßen beschrieben: „Wenn in einem Dokument top_1 sichtbar ist, so ist in diesem Dokument (zu einem gewissen Grad/mit einer gewissen Wahrscheinlichkeit) auch top_2 sichtbar.“ Das heißt in diesem Fall: Zu einem Grad von 0,2 bzw. mit einer Wahrscheinlichkeit von 20% behandelt eine Webseite, die top_1 enthält, auch top_2 (ohne dass notwendigerweise der String top_2 in der Webseite vorkommen muss). Da es hier nun 100 Seiten mit top_1 gibt, ist es nahe liegend, 20 Seiten davon (= 20%) bei top_2 mitzuzählen, während sich die Sichtbarkeit von top_1 nicht verändert, da es keine eingehenden Kanten hat. Hätte top_2 zwei Inputkanten wie in Abbildung 32, so würde sich der neue Wert für top_2 auf die gleiche Weise aus der Summe seines alten Wertes und aller eingehenden Kanten bilden.

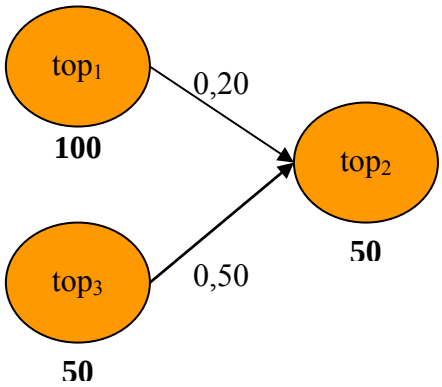


Abbildung 32: Themennetzwerk mit drei Knoten

Allgemein formuliert ergibt sich[1]:

$$\text{visibility}_1(\text{top}) = \text{visibility}_0(\text{top}) + \sum_{x \in TOP} \text{visibility}_0(x) \cdot \text{topicrelationships}(x, \text{top})$$

und gleichzeitig gilt visibility’(top) = visibility_1(top), d.h. der Algorithmus terminiert nach einer Runde – der Endwert ist gleich dem Wert nach einem Propagierungsschritt. Dies verstößt jedoch gegen Forderung (2) an den Algorithmus (vgl. 5.), nämlich die Unterstützung transitiver Beziehungen (vgl. Abbildung 33): Hier würden zwar jeweils top_1 top_2 und top_2 top_3 beeinflussen, nicht jedoch die Sichtbarkeit von top_1 bis auf top_3 ausstrahlen. Es bietet sich eine zweimalige Durchführung des Algorithmus’ an und es ist offensichtlich, dass die Anzahl

[1] Beachte: topicrelationships(top_a, top_a) = 0

der notwendigen Schritte allgemein dem Diameter des Graphen entspricht, d.h. visibility'(top) = $visibility_{Diameter}(top)$

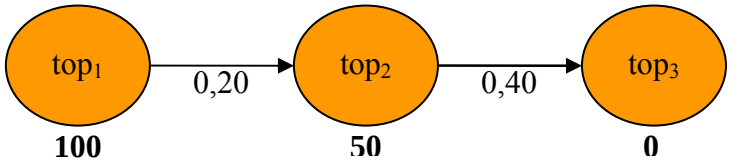


Abbildung 33: Themennetzwerk mit drei Knoten (transitive Beziehung)

t	$visibility_t(top_1)$	$visibility_t(top_2)$	$visibility_t(top_3)$
0	100	50	0
1	100	70	20
2 = Diameter	100	90	40

Tabelle 6: Ergebnis von SimplePropagation nach zwei Schritten

Tabelle 6 zeigt das Ergebnis für SimplePropagation nach zwei Schritten (= Diameter) für das Themennetzwerk aus Abbildung 33: Die Werte scheinen intuitiv betrachtet etwas zu hoch, doch gerade noch im Rahmen; deutlich wird das Versagen, wenn man einen höheren Diameter unterstellt, wenn also rechts von top_3 auf die gleiche Weise noch weitere Knoten angehängt würden: In so einem Fall liefen die Sichtbarkeiten für top_2 völlig aus dem Ruder, denn pro Runde würde die Sichtbarkeit von top_2 um $100 \cdot 0{,}2 = 20$ zunehmen, was nach 100 Runden einen sicherlich unsinnigen Wert von 2050 ergäbe; und das bei einem mit 0,2 sehr niedrigen Gewicht zwischen top_1 und top_2.

Schon aus diesem Grund wird SimplePropagation an dieser Stelle sofort verworfen. Die Ausführungen hierzu waren wie gesagt trivial, sollten jedoch eine Einführung in die Überlegungen zu den folgenden Algorithmen bieten.

5.2.2 DeltaPropagation

Eine Möglichkeit, die Zahlen nicht ins Unermessliche steigen zu lassen, besteht darin, nur den Unterschied der Sichtbarkeitswerte zwischen zwei Themen zu propagieren, d.h.

$$visibility_t(top) = visibility_{t-1}(top) + \sum_{x \in TOP} (visibility_{t-1}(x) - visibility_{t-1}(top)) \cdot topicrelationships(x, top)$$

Tabelle 7 zeigt das Ergebnis für DeltaPropagation für das Themennetzwerk aus Abbildung 33 nach zwei Schritten (= Diameter): Die Werte sind günstiger als die des SimplePropagation-Algorithmus' aus Tabelle 6, doch es ergibt sich abermals ein Problem, wenn man einen höheren Diameter unterstellt: Schon nach 20 Schritten streben alle drei Werte auf den gleichen Wert 100 zu (was zu erwarten war).

t	$visibility_t(top_1)$	$visibility_t(top_2)$	$visibility_t(top_3)$
0	100	50	0
1	100	60	20
2	100	68	36
20	100	99,42	98,84

Tabelle 7: Ergebnis von DeltaPropagation nach zwei und nach 20 Schritten

Dem aufmerksamen Leser ist sicherlich weiterhin aufgefallen, dass in obenstehender Formel und auch in diesem Beispiel nur solche Fälle behandelt wurden, in denen die Sichtbarkeit des Themas, bei dem die Kante beginnt, höher ist als diejenige, bei der die Kante endet. Was jedoch, wenn dies nicht der Fall ist? Eine Abnahme der Sichtbarkeiten wurde durch Forderung (1) ausgeschlossen, so dass die Formel so nicht einsetzbar ist. Eine Ergänzung um Betragsstriche ist auch nicht möglich, da sonst eine Konstellation 99 → 100 einen kleineren Einfluss hätte als 50 → 100. Eine weitere Möglichkeit wäre, in diesem Fall einfach gar keinen Sichtbarkeitsausgleich durchzuführen, d.h. nur Kanten zu berücksichtigen, bei denen das „linke" Thema sichtbarer ist als das „rechte". Dies hätte zur Folge, dass man sich die Kante bei der Modellierung hätte sparen können (vgl. Abbildung 34), was auch nicht erwünscht ist.

Abbildung 34: Themennetzwerke mit und ohne Kante entsprechen sich bei DeltaPropagation

Auch DeltaPropagation wird aus diesen beiden Gründen verworfen.

5.2.3 Propagation mit Decay

Der nächste Versuch, die ins Unendliche strebenden Werte der SimplePropagation unter Kontrolle zu bekommen, ist die Anwendung eines Decay ($d \in [0..1]$), d.h. die Gewichte werden in jeder Runde etwas kleiner ($g_t = g_{t-1} * d$) und streben gegen 0, so dass sich auch die entlang der Kanten propagierten Werte in jeder Runde verringern bis die Sichtbarkeitswerte schließlich konvergieren.

$$\text{visibility}_t(\text{top}) = \text{visibility}_{t-1}(\text{top}) + \sum_{x \in TOP} \text{visibility}_{t-1}(x) \cdot \text{topicrelationships}(x, \text{top}) \cdot d^t$$

Angewandt auf das Themennetzwerk aus Abbildung 33 ergeben sich für d = 0,5 folgende Werte (vgl. Tabelle 8):

t	$\text{visibility}_t(\text{top}_1)$	$\text{visibility}_t(\text{top}_2)$	$\text{visibility}_t(\text{top}_3)$
0	100	50	0
1	100	70	20
2	100	80	34
20	100	90,00	50,67

Tabelle 8: Ergebnis von Propagation mit Decay nach zwei und 20 Schritten

Die Sichtbarkeit von Thema 2 strebt gegen 90, die von Thema 3 gegen 50,67 (zum Zeitpunkt t=20 war Konvergenz sichtbar gegeben). Diese Werte erscheinen etwas hoch, doch immerhin konvergiert das Netz. Daher soll dieser Algorithmus – im Folgenden mit (DEC) bezeichnet – prinzipiell in Betracht gezogen werden und in Abschnitt 7 im Rahmen der Evaluierung mit der anderen Möglichkeit – einem auf Spreading Activation basierenden Algorithmus – verglichen werden.

5.2.4 Spreading Activation

5.2.4.1 Anpassung des ursprünglichen Spreading Activation Algorithmus'

Der in dieser Arbeit präferierte Algorithmus zum Sichtbarkeitsausgleich basiert auf der in 5.1.2 vorgestellten Spreading Activation, die für diese Anwendung zunächst aus folgendem Grund angepasst werden muss:

Spreading Activation geht von einem Energiestoß aus, der sich entlang der Kanten auf die Nachbarknoten ausbreitet, aber dabei in der Summe der Gesamtenergie nicht mehr und nicht weniger wird. Dies basiert auf der physikalischen Vorstellung von Energie, die weder verloren gehen noch aus dem Nichts kommen kann. In Abbildung 35 beispielsweise entsprechen sich die semantischen Netze (I) und (II): In beiden Fällen wird der Energiestoß, der aus dem schraffierten Knoten kommt, proportional zu den Kantengewichten auf die drei Nachbarknoten verteilt, d.h. in beiden Fällen wird die Energie im Verhältnis 1 : 1 : 2 weitergeleitet.

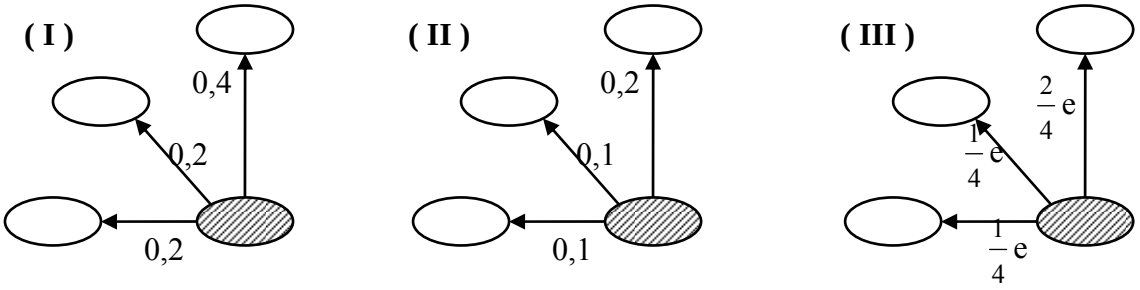


Abbildung 35: Proportionalisierung der Energie in Spreading Activation

Im einfachsten Fall bedeutet dies sogar, dass ein Knoten, der nur einen Nachbarknoten hat (vgl. z.B. Abbildung 12), in jedem Fall seine gesamte Energie an diesen weiterleitet – unabhängig von der Höhe des zwischen den Knoten bestehenden Gewichts. Gerade dies ist beim Sichtbarkeitsausgleich undenkbar, denn zwei Themen, die in einer engen Beziehung stehen und daher ein hohes Gewicht haben, sollen mehr Sichtbarkeit zwischen sich austauschen als Themen, die mit einer schwach gewichteten Kante verbunden sind. Anders gesagt bedeutet die Tatsache, dass top_1 nur einen Nachbarn hat, nicht, dass alle Seiten, die top_1 enthalten, die Sichtbarkeit von top_2 erhöhen – dies darf nicht von der Anzahl der Nachbarn abhängen, sondern von der Enge der bestehenden Beziehungen.

Außerdem entspricht die Vorstellung, dass die Gesamtsumme der Sichtbarkeiten (wobei Sichtbarkeit hier der Energie entspricht) sich nicht verändern darf, nicht der bisher in dieser Arbeit vertretenen Sichtweise auf Sichtbarkeiten.

Die aus diesem Grund notwendige Anpassung wird sichtbar in dem folgenden mit (SA) bezeichneten Algorithmus – einer Änderung des Algorithmus' aus 5.1.2: Beim rekursiven Aufruf von energize auf die Nachbarknoten wird nicht das zur Proportionalisierung vorher berechnete e' mit dem Gewicht multipliziert, sondern direkt die Gesamtenergie des Energiestoßes e verwendet:

(SA)

```
procedure energize (e ∈ R+0 , t ∈ TOP) {
    visibility(t) ← visibility(t) + e;
    if e > T then
        ∀n ∈ TOP:
            if topicrelationships(t, n) > 0 then
                energize (e · topicrelationships(t, n), n);
            end if
    end if
}
```

Die Funktion der verwendeten Parameter wurde bereits in 5.2.1 erklärt und hier auf die in dieser Arbeit eingeführten Definitionen angepasst.

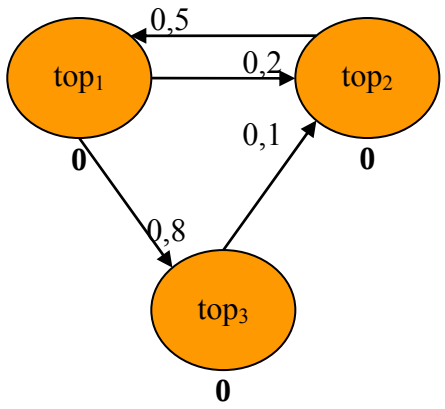
Zyklische Beziehungen in Spreading Activation bewirken, dass im Laufe eines Energiestoßes die Energie immer im Kreis herumgereicht und dabei kleiner wird; für diesen Fall dient der Parameter T als Abbruchkriterium, der so klein gewählt werden sollte, dass die Werteveränderungen minimal genug sind, um von Konvergenz sprechen zu können.

Beim Sichtbarkeitsausgleich wird Energie nicht nur in einem einzelnen Knoten zugeführt, sondern in allen gleichzeitig, da initiale Sichtbarkeit in allen Knoten vorhanden ist. Umzusetzen ist dies folgendermaßen: Die Knotengewichte werden zunächst auf 0 gesetzt und anschließend der Algorithmus nacheinander für jedes der Themen angewandt. Daraus ergibt

sich am Ende für jedes Knotengewicht die Summe der eigenen Aktivierung mit den Energieteilen, die im Rahmen der Aktivierung jedes anderen Themas von Nachbarknoten zugeflossen sind. Im folgenden Abschnitt wird dies an einem kleinen Beispiel gezeigt.

5.2.4.2 Ein Beispiel zur Berechnung

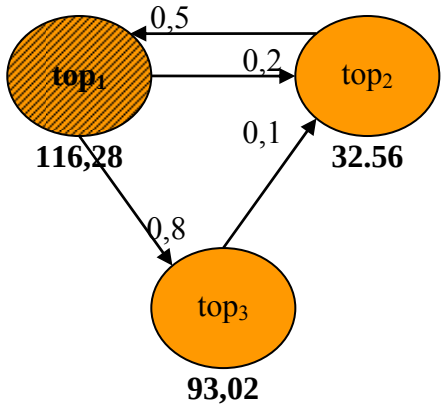
Abbildung 36 zeigt ein Themennetzwerk mit drei Knoten und vier Beziehungen, wobei auch Zyklen vorhanden sind. Die initialen Sichtbarkeiten sind hier nicht unter den Knoten als Knotengewichte dargestellt, sondern in nebenstehender Tabelle. Diese Sichtbarkeiten werden nun nacheinander in ihren jeweiligen Knoten als Energiestoß im Sinne der Spreading Activation injiziert; es soll mit Thema 1 begonnen werden.



i	$visibility_0(top_i)$
1	100
2	50
3	10

Abbildung 36: Beispiel zur Berechnung von Spreading Activation; initiales Netz

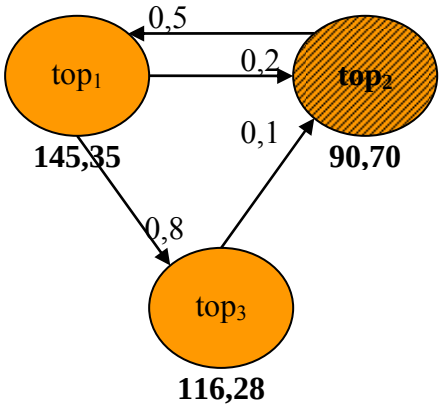
Die Sichtbarkeiten nach Aktivierung von Thema 1 zeigt Abbildung 37: Thema 1 hat als in diesem Schritt aktivierter Knoten am meisten zugewonnen und zwar mehr als die ihm zugeführte initiale Sichtbarkeit von 100: Das liegt daran, dass er über die Zyklen top_1 → top_3 → top_2 → top_1 sowie top_1 → top_2 → top_1 Teile seiner Aktivierungsenergie wieder zurück erhält. top_3 hat mit 93,02 ebenfalls einen sehr hohen Wert, was durch die hohe Gewichtung der Kante top_1 → top_3 zu erklären ist.



i	Δe bei Aktivierung von		
	top1	top2	top3
1	**116,28**		
2	**32,56**		
3	**93,02**		

Abbildung 37: Beispiel zur Berechnung von Spreading Activation; nach der ersten Aktivierung

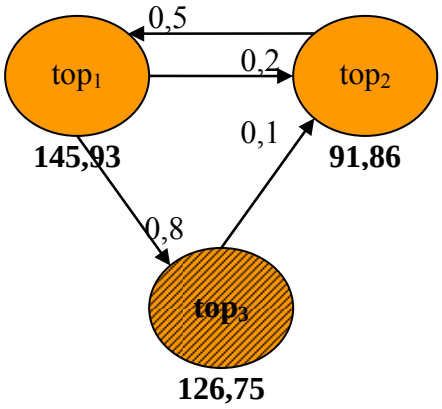
Die Werte nach dem zweiten Schritt zeigt Abbildung 38:



i	Δe bei Aktivierung von		
	top1	**top2**	top3
1	116,28	**29,07**	
2	32,56	**58,14**	
3	93,02	**23,26**	

Abbildung 38: Beispiel zur Berechnung von Spreading Activation; nach der zweiten Aktivierung

Die Zahlen unter den Knoten sind in diesem Fall bereits die aufsummierten Werte der beiden Aktivierungsschritte. Entsprechend zu oben wird nun die dritte und letzte Aktivierung durchgeführt und man erhält das folgende Ergebnis (vgl. Abbildung 39):



i	Δe bei Aktivierung von		
	top1	top2	**top3**
1	116,28	29,07	**0,58**
2	32,56	58,14	**1,16**
3	93,02	23,26	**10,47**

Abbildung 39: Beispiel zur Berechnung von Spreading Activation; nach der dritten Aktivierung

Thema	initiale Sichtbarkeit	relative intitiale Sichtbarkeit	abgeleitete Sichtbarkeit	relative abgeleitete Sichtbarkeit
top_1	100	62,50 %	145,93	40,03 %
top_2	50	31,25 %	92,86	25,20 %
top_3	10	6,25 %	126,75	34,77 %
Summe	160	100,00 %	365,54	100,00 %

Tabelle 9: Beispiel zur Berechnung von Spreading Activation, Endergebnisse

Tabelle 9 listet die initialen und abgeleiteten Sichtbarkeiten dieses Beispiels (jeweils in absoluter und relativer Form) auf. Je nach Anwendungsfall kann sich der Anwender für die relativen oder absoluten Sichtbarkeiten interessieren.

5.2.4.3 Diskussion des gewählten Algorithmus'

Die nun folgenden Überlegungen sollen zeigen, dass dieser Algorithmus von der intuitiven Vorstellung her dem Sichtbarkeitsausgleich am ehesten angemessen ist. Ruft man sich das Vorgehen der naiven SimplePropagation (vgl. 5.2.1) in Erinnerung, so wird dort das Problem der transitiven bzw. zyklischen Beziehungen derart gelöst, dass einfach derselbe Schritt mehrfach hintereinander durchgeführt wird, bis man sicher sein kann, dass alle Winkel des Themennetzwerks erreicht worden sind. An einem einfachen Beispiel (vgl. Abbildung 33, Tabelle 6) ergaben sich dadurch unrealistisch hohe Werte für top_2 (2050 bei Diameter 20), da die initiale Sichtbarkeit von top_1 in Höhe von 100 in jedem der Schritte an top_2 weitergereicht wurde. Diese mehrfache Berücksichtigung der gleichen 100 läuft der Vorstellung eines Sichtbarkeitsausgleichs stark zuwider: Interpretiert man die 100 gemäß der in dieser Arbeit gängigen Vorstellung von Sichtbarkeitswerten als Anzahl der Webseiten und die Kante von 0,2 als Wahrscheinlichkeit bzw. Anteil von Seiten, so ist es durch nichts zu rechtfertigen, warum plötzlich $100 \cdot 0{,}2 \cdot t$ Seiten bei top_2 als relevant mitgezählt werden sollen (anstatt der durch die Kante ausgedrückten $100 \cdot 0{,}2$). Anders gesagt: Der Umgang mit der Transitivität betrifft lediglich den Teil der Sichtbarkeit, der im ersten Schritt von top_1 zu top_2 weitergereicht wurde (= 20), nicht jedoch die ursprünglichen 100. Es wäre also beim Beispiel aus Abbildung 33 bei weitem sinnvoller, im zweiten Schritt die Beziehung zwischen top_1 und top_2 nicht mehr zu

berücksichtigen und sich nur noch um die von top_2 neu hinzugewonnene Sichtbarkeit von 20 zu kümmern.
Doch dies ist genau das Vorgehen der Spreading Activation, die diese 20 als Energiepaket interpretiert, das als Aktivierungswelle durch das Netzwerk fließt und bei einem Netzwerk ohne Zyklen wie in Abbildung 33 schließlich verebbt.
Auch die Propagation mit Decay berücksichtigt den gleichen Sichtbarkeitswert von 100 mehrfach, doch durch die abnehmenden Gewichte in jedem Schritt etwas weniger. Trotz der dadurch erreichten Konvergenz widerspricht diese mehrfache Einrechnung des gleichen Wertes der intuitiven Idee eines Sichtbarkeitsausgleichs.

Schließlich bleibt noch der Abgleich mit den in 3.2 formulierten Anforderungen:
(1) ist auf jeden Fall gegeben, da die Sichtbarkeiten nicht abnehmen können. Jedes Thema hat durch die erste Zeile der energize-Prozedur mindestens seinen Startwert auch als Endwert (vgl. auch das Beispiel in 5.2.4.2). Transitive Beziehungen (Anforderung 2) werden – wie oben dargelegt – unterstützt. (3) und (4), die Stabilität gegenüber kleinen Veränderungen der Kantengewichte bzw. der Anfangswerte, sollen in der Evaluierung in Abschnitt 7 behandelt werden. Die Ergebnisse lagen bei dem Beispiel in 5.2.4.2 im Rahmen des Sinnvollen (Anforderung 6) und die mit Sicherheit auftretende Konvergenz verhindert in jedem Fall ein Streben der Werte gegen unendlich, doch auch dies soll an weiteren Beispielen in der Evaluierung getestet werden.
Der einzige noch fehlende Punkt (5) ist die Berücksichtigung von Covibilitätswerten, dem sich der nun folgende Abschnitt widmet.

5.3 Berücksichtigung von Covisibilitäten

5.3.1 Grundlegende Überlegungen

Bevor die Berücksichtigung der Covisibilitäten (=Co-Sichtbarkeiten aus Punkt 2.3) in einen konkreten Algorithmus integriert wird, sollen zunächst grundlegende Überlegungen zu Covisibilitäten im Rahmen des Sichtbarkeitsausgleichs angestellt werden.

Man stelle sich zwei in sehr enger Beziehung zueinander stehende Themen vor – ein gutes Beispiel sind „HIV“ und „Aids“: Das HI-Virus löst das Krankheitsbild Aids aus und Aids kann ausschließlich durch HIV entstehen, so dass sie in einer wechselseitigen Assoziationsbeziehung mit hohen Kantengewichten stehen. Im Themennetzwerk in Abbildung 13 besteht eine mit 0,70 gewichtete Kante von Aids nach HIV und eine 0,85-Kante in die Gegenrichtung, in Abbildung 40 noch mal als Teilausschnitt mit anderen initialen Sichtbarkeiten:

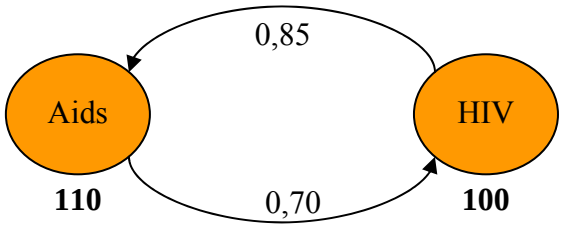


Abbildung 40: Themennetzwerk mit zwei zyklisch verbundenen Knoten, Aids und HIV

Auf dieses kleine Themennetzwerk werde der SimplePropagation-Algorithmus mit nur einem Schritt angewandt. Die einfach zu errechnenden Ergebnisse sind in Tabelle 10 gelistet:

Thema	$visibility_0$	visibility'
Aids	110	195
HIV	100	177

Tabelle 10: Sichtbarkeiten der Themen aus Abbildung 40 nach einem Schritt SimplePropagation ohne Berücksichtigung der Covisibilitäten

Bei hohen Kantengewichten wie 0,7 und 0,85 erstaunt der starke Anstieg der Sichtbarkeitswerte kaum. Das ist an sich korrekt und einsichtig, solange man davon ausgeht, dass die Covisibilität der beiden Themen 0 ist. Sollte jedoch (im hier mit Absicht skizzierten Extremfall) auf sämtlichen 100 HIV-Seiten auch der Begriff „Aids“ auftauchen[1], so wären diese 100 HIV-Seiten bei den 110 Aids-Seiten bereits mitgezählt. Mit den Definitionen für Covisibilitätsmaße aus 2.3.2.1 ergäben sich die Covisibilitätswerte aus Tabelle 11[2].

[1] Hier wird wiederum am Bild der Webseiten argumentiert; eine Interpretation der Covisibilität in anderen internetbasierten Kommunikationsprozessen müsste entsprechend formuliert werden.

[2] Vgl. hierzu auch Tabelle 1: Die realen Covisibilitäten zwischen „Aids“ und „HIV“.

Ist es nun wünschenswert, dass HIV-Seiten, die bereits durch die Suchmaschine bei „Aids“ mitgezählt wurden, durch den Algorithmus (sei es SimplePropagation, Spreading Activation oder ein anderer) abermals von „HIV“ nach „Aids“ propagiert werden? Diese Frage ist sicherlich mit „nein“ zu beantworten, so dass eine Anpassung des Algorithmus’ dringend erforderlich ist.

Trefferzahl HIV	100
Trefferzahl Aids	110
Trefferzahl HIV AND Aids	100
covisibility(HIV, Aids)	0,91
$covisibility_1$(HIV, Aids)	1
$covisibility_2$(HIV, Aids)	0,91

Tabelle 11: Beispielhaft hohe Co-Visibilitäten der Themen „Aids“ und „HIV“

Hierzu soll zunächst wieder ein einzelner SimplePropagation-Schritt betrachtet werden: Bei der Propagierung der Sichtbarkeit sollen nur diejenigen Seiten berücksichtigt werden, die lediglich „HIV“, nicht jedoch „Aids“ enthalten: Aus Tabelle 11 geht schnell hervor, dass dies in diesem Fall 0 Seiten sind.
Es ist einsichtig, dass also zur Berechnung der zu berücksichtigenden Seiten das Covisibilitätsmaß $covisibility_1$ hinzugezogen werden muss, denn:

$covisibility_1(x, y)$	ist als der Anteil der Seiten mit „x AND y“ an der Zahl der Seiten, die nur „x“ enthalten definiert.
$1 - covisibility_1(x, y)$	ist demnach der Anteil der Seiten mit „x AND NOT y“ an der Zahl der Seiten, die nur „x“ enthalten, was genau der gesuchte Prozentsatz ist.

Daraus entsteht folgende veränderte Propagierungsformel:

$$visibility_t(top) = visibility_{t-1}(top) + \sum_{x \in TOP} visibility_{t-1}(x) \cdot (1 - covisibility_1(x, top)) \cdot topicrelationships(x, top)$$

Berücksichtigt man auf diese Weise die Covisibilitäten aus Tabelle 11 im Themennetzwerk aus Abbildung 40, so entstehen – im Kontrast zu den Ergebnissen ohne Covisibilität aus Tabelle 10 – die in Tabelle 12 aufgeführten Ergebnisse, die kaum eine Veränderung zu den initialen Sichtbarkeiten darstellen. Dies ist auf Grund der hier unterstellten extrem hohen Covisibilität durchaus sinnvoll. Die Sichtbarkeit von „Aids“ hat sich sogar gar nicht verändert, da die 100 HIV-Seiten bereits komplett von der Suchmaschine bei „Aids“ mitgezählt wurden und somit im Algorithmus keine neue Sichtbarkeit für „Aids“ schaffen.

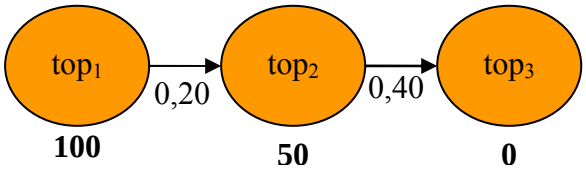
Thema	**$visibility_0$**	**visibility'**
Aids	110	110
HIV	100	106,93

Tabelle 12: Sichtbarkeiten der Themen aus Abbildung 40 nach einem Schritt SimplePropagation mit Berücksichtigung der Covisibilitäten

Allgemein weisen eng verwandte Themen wie die hier gewählten „Aids“ und „HIV“ sicherlich mit großer Wahrscheinlichkeit eine hohe Covisibilität auf (vgl. 4.3.2.2) – wenn auch nicht so hoch wie soeben im Extremfall unterstellt. Das heißt interessanterweise, dass Themen mit hoch gewichteten Kanten zwischen sich oft auch eine hohe Covisibilität haben und somit bei Anwendung oben stehender Formel der Effekt hoher Gewichte durch die bremsende Wirkung der Covisibilität konterkariert wird.

5.3.2 Covisibilitäten im Spreading Activation Algorithmus

Doch wie sind die Covisibilitäten bei der weiteren Propagierung im zweiten und allen weiteren Schritten zur Berücksichtigung der transitiven Beziehungen zu handhaben? Hierzu soll wieder das transitive Beispiel aus Abbildung 33 betrachtet werden, ergänzt um Covisibilitäten (vgl. Abbildung 41):



covisibility$_1$(top$_i$, top$_j$)				
i	j →	1	2	3
1		1	0,5	0,4
2		0,1	1	0,6
3		0,2	0,1	1

Abbildung 41: Themennetzwerk mit drei Knoten (transitiv) und Covisibilitäten

Man beachte hierbei, dass stets covisibility$_1$(A, A) = 1 gilt, was nach der Definition der covisibility$_1$ trivialerweise gilt, denn der Anteil der Seiten, die A und A enthalten, an den Seiten, die A enthalten, ist stets 100%.

In Abbildung 41 wird ein Propagierungsschritt mit der Formel aus 5.3.1 durchgeführt und es ergeben sich die folgenden Werte:

$$\text{visibility}_1(\text{top}_1) = 100$$
$$\text{visibility}_1(\text{top}_2) = 50 + 100 \cdot 0{,}2 \cdot 0{,}5 = 50 + 10 = 60$$
$$\text{visibility}_1(\text{top}_3) = 0 + 50 \cdot 0{,}4 \cdot 0{,}4 = 0 + 8 = 8$$

Der Sichtbarkeitszuwachs beträgt also 10 bei top$_2$ und 8 bei top$_3$. Spreading Activation verwendet im nächsten Schritt genau diese Zuwächse und gibt einen Teil davon an die jeweils nächsten Nachbarknoten weiter. top$_3$ hat keine folgenden Nachbarknoten, so dass die 8 sich an dieser Stelle nicht weiterverbreiten, top$_2$ jedoch gibt einen Teil seines Zuwachses von 10 an top$_3$ weiter. Doch welche Covisibilität ist hierbei zu berücksichtigen? Die zwischen top$_2$ und top$_3$ (denn dies ist die Kante, an der im Moment propagiert wird) oder diejenige zwischen top$_1$ und top$_3$ (denn von top$_1$ kam die propagierte Sichtbarkeit ursprünglich)?

(A)

$$\text{visibility}_2(\text{top}_3) = \text{visibility}_1(\text{top}_3) + \text{visibility}_1(\text{top}_2) \cdot \text{topicrel}\,(\text{top}_2, \text{top}_3) \cdot \mathbf{covis_1(top_2, top_3)}$$

ODER

(B)

$$\text{visibility}_2(\text{top}_3) = \text{visibility}_1(\text{top}_3) + \text{visibility}_1(\text{top}_2) \cdot \text{topicrel}\,(\text{top}_2, \text{top}_3) \cdot \mathbf{covis_1(top_1, top_3)}$$

Beide Versionen sind gleichermaßen denkbar: Version (B) argumentiert eher, dass Transitivität einem Ausstrahlen der Sichtbarkeit von top_1 bis auf top_3 gleichkommt und somit der Ursprungsknoten noch relevant ist, Version (A) folgt eher der Vorstellung, dass die im ersten Propagierungsschritt von top_2 akkumulierte Sichtbarkeit voll in den „Besitz" von top_2 übergegangen ist und der Ursprung nicht mehr entscheidend ist. Technisch gesehen entspricht (A) lediglich einer Veränderung der Kantengewichte: Die vom Nutzer bei der semantischen Modellierung festgelegten Kantengewichte werden durch die Covisibilitätswerte vermindert. Für eine Implementierung von Version (B) hingegen ist zu beachten, dass zu jedem propagierten Sichtbarkeitspaket der Ursprungsknoten mitgespeichert werden muss. Diese beiden abgeänderten Versionen des Spreading Activation Algorithmus' (SA) aus 5.2.4.1 sollen im Folgenden (SA-A) und (SA-B) genannt werden:

(SA-A)

```
procedure energize (e ∈ R+0 , t ∈ TOP) {
    visibility(t) ← visibility(t) + e;
    if e > T then
        ∀n ∈ TOP:
            if topicrelationships(t, n) > 0 then
                energize (e · topicrelationships(t, n) · (1-covis(t, n)), n);
            end if
    end if
}
```

(SA-B)

```
procedure energize (e ∈ R+0 , t ∈ TOP, source ∈ TOP) {
    visibility(t) ← visibility(t) + e;
    if e > 0 then
        ∀n ∈ TOP:
            if topicrelationships(t, n) > 0 then
                energize (e · topicrelationships(t, n) · (1-covis(source, n)),
```

```
                                    n, source);
                end if
        end if
    }
```

covis(t, n) wurde hierbei als Abkürzung für $covisibility_1(t, n)$ verwendet. Der Parameter source in (SA-B) wird bei jedem rekursiven Prozeduraufruf weitergereicht und ist der Knoten, von dem der Energiestoß ausging. Bei erstmaligem Aufruf der Prozedur gilt also t = source.

Eine interessante Auswirkung zeigt (SA-B) bei der Anwendung auf Zyklen: In dem Moment, in dem ein Energie-(Sichtbarkeits-)Paket im Kreis herumgereicht wurde und wieder bei seinem Ursprungsknoten ankommt, gilt covis(a, a) = 1 und somit (1-covis(a,a)) = 0, so dass an dieser Stelle ein Energiepaket der Größe 0 propagiert wird. Daher ist T in (SA-B) konstant als 0 gewählt: Es wird erst abgebrochen, sobald an einer Kante ein leeres Energiepaket propagiert werden soll, d.h. wenn die Covisibilität 1 war. Ein Thema kann sich in (SA-B) demnach nicht über Zyklen selbst beeinflussen, d.h. der Algorithmus (SA-B) konvergiert nicht dadurch, dass die Energie irgendwann unter eine Schranke T fällt, sondern endet nach Durchlaufen des gesamten Netzes von selbst.

Ob eine Berücksichtigung der Covisibilitäten sinnvoll ist, ist letztendlich nur an praktischen Beispielen zu beantworten. Führt die Anwendung von (SA-A) oder (SA-B) zu signifikant anderen Ergebnissen als die Anwendung des einfachen (SA)? Sind reale Covisibilitäten überhaupt hoch genug, um einen Einfluss zu haben? Dieser Frage soll in Punkt 7.4.3.4 im Rahmen der Evaluierung nachgegangen werden.

6. SOFTWARETECHNISCHE UMSETZUNG

6.1 Überblick

Ziele der Realisierung im Rahmen dieser Diplomarbeit waren vor allem zweierlei:

1. Schaffung eines Tools zum Testen der Algorithmen in verschiedenen Themennetzwerken.
2. Schaffung eines in jedem Kontext verwendbaren Werkzeugs zum Beobachten von Themen im Internet an Hand ihres Recall-Wertes.

Entgegen des in dieser Arbeit vertretenen Ansatzes wurde das Konzept des Thesaurus mit Ableitung des Themennetzwerks nicht umgesetzt, sondern eine direkte Modellierung des Themennetzwerks vorgezogen. Einerseits bergen weder die Programmierung eines Thesaurus-Editors noch die nutzergestützte Ableitung des Themennetzwerks interessante Aspekte in der Umsetzung, und andererseits ist der Zeitaufwand der GUI-Programmierung für einen solchen Thesaurus-Editor extrem hoch. Daher wird im Rahmen der softwaretechnischen Realisierung davon ausgegangen, dass nur überschaubare Themennetzwerke modelliert werden (d.h. dass eine kognitive Unterstützung durch Thesaurus nicht unbedingt notwendig ist) und dass es sich weiterhin um erfahrene Nutzer handelt, die im Stande sind, direkt ein korrektes Themennetzwerk ohne den Umweg über den Thesaurus zu kreieren.

Die softwaretechnische Umsetzung erfolgte in Java (Java 2 Standard Edition, JDK 1.4). Die GUI basiert auf Swing, wobei die Opensource JGraph-API (http://www.jgraph.com/) zur Darstellung der Themennetzwerke verwendet wurde. Zur Ermittlung der Recall-Werte dienten die Google™-WebAPIs.
Im Folgenden soll in Punkt 6.2 der Aufbau des Programms an Hand der in ihm verwendeten Dokumenttypen und ihrer Verwendung in einem typischen Arbeitsablauf skizziert werden

sowie in Punkt 6.3 ein Erfahrungsbericht zum Einsatz der Google™-WebAPIs gegeben werden. Auf weitere Ausführungen zur softwaretechnischen Realisierung wird verzichtet, da die Umsetzung der Algorithmen oder der Entwurf der notwendigen Klassen keinerlei interessanten Aspekte bietet.

6.2 Dokumenttypen

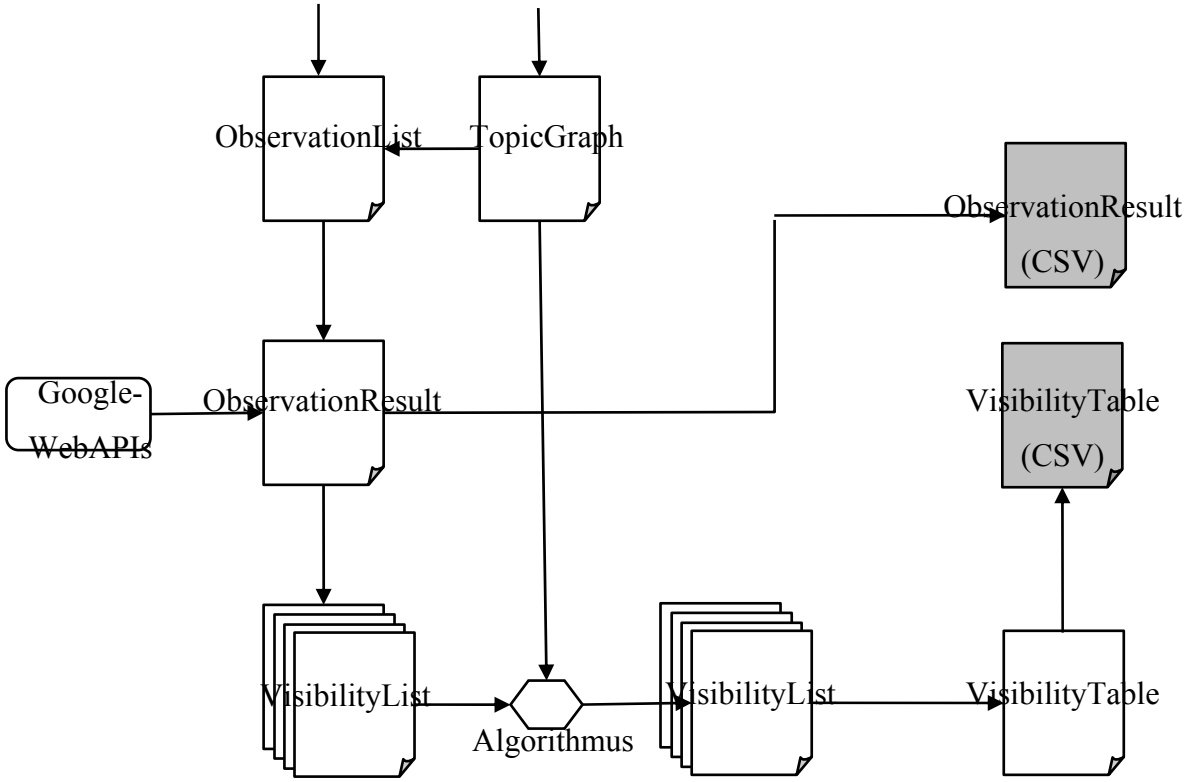


Abbildung 42: Zusammenhang zwischen den verschiedenen Dokumenttypen

Alle Dokumenttypen sind in ihrer persistenten Form als XML-Dateien abspeicherbar.

Ausgangspunkt ist die Modellierung eines *TopicGraph* (entspricht dem Themennetzwerk) durch den Nutzer. Ein TopicGraph enthält Themen als Knoten mit Bezeichnungen sowie die Beziehungen zwischen den Themen als gewichtete Kanten (vgl. Abbildung 43).

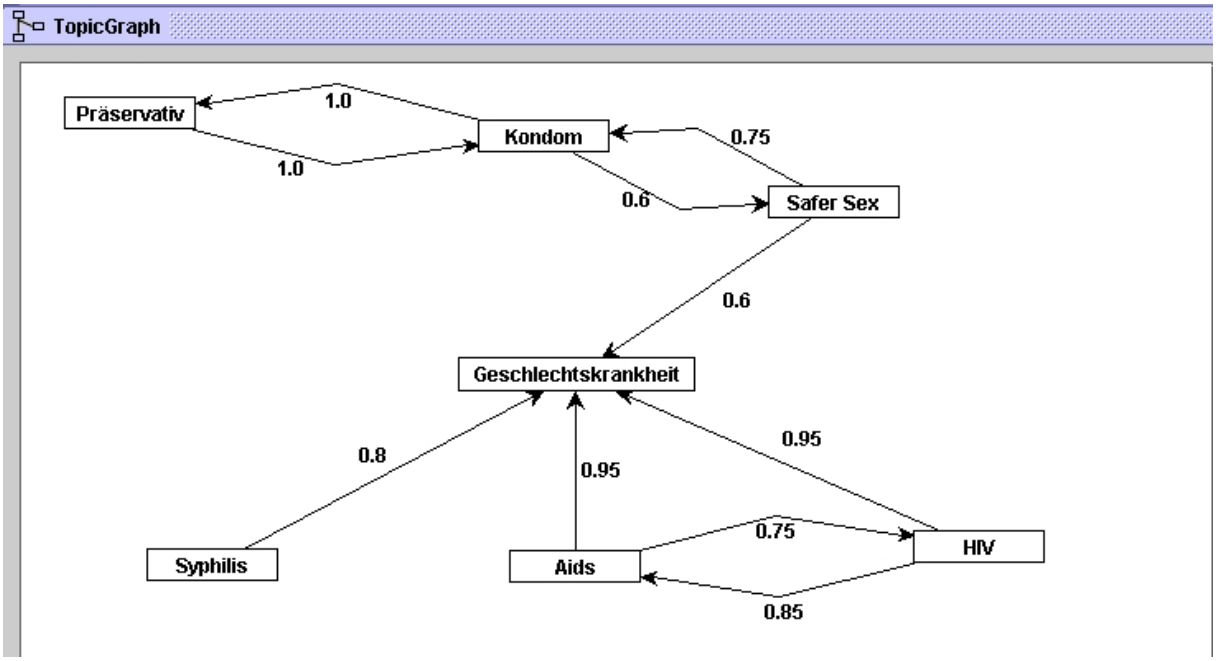


Abbildung 43: TopicGraph

Eine *ObservationList* speichert eine beliebige Anzahl an Begriffen (Strings) und Domains (Strings), in denen diese Begriffe beobachtet werden sollen (vgl. Abbildung 44). Es ist möglich, eine ObservationList direkt zu erstellen oder aus einem TopicGraph abzuleiten. Bei letzterem Vorgehen wird aus der Bezeichnung jedes Themas im TopicGraph ein Begriff in der ObservationList. Zusätzlich ist es möglich, die Covisibilitäten zwischen den Themen in die ObservationList als mit „AND“ verknüpfte Begriffe aufzunehmen.

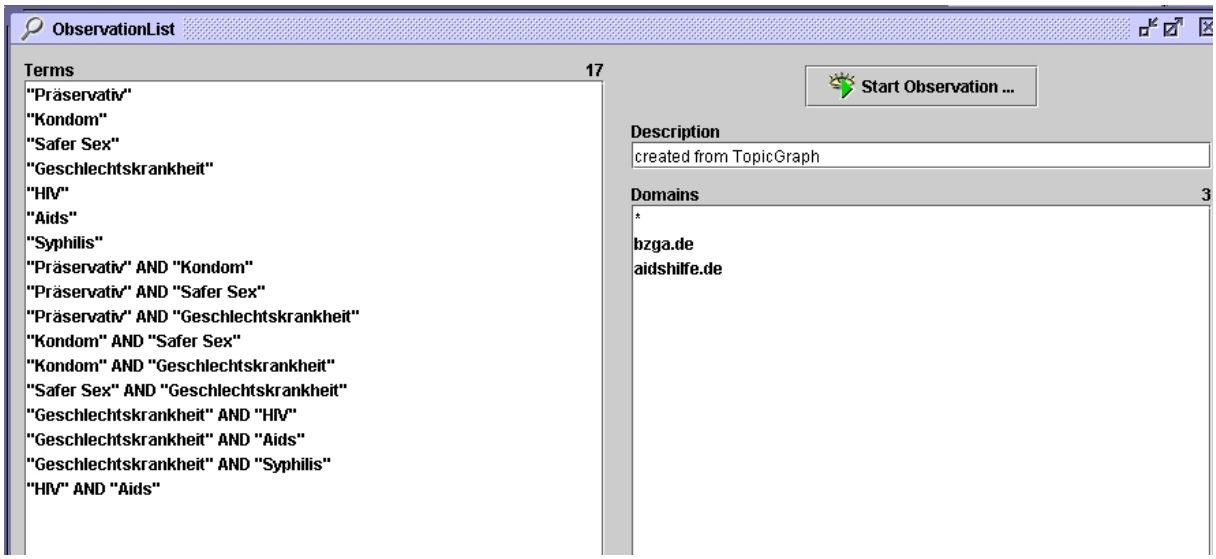
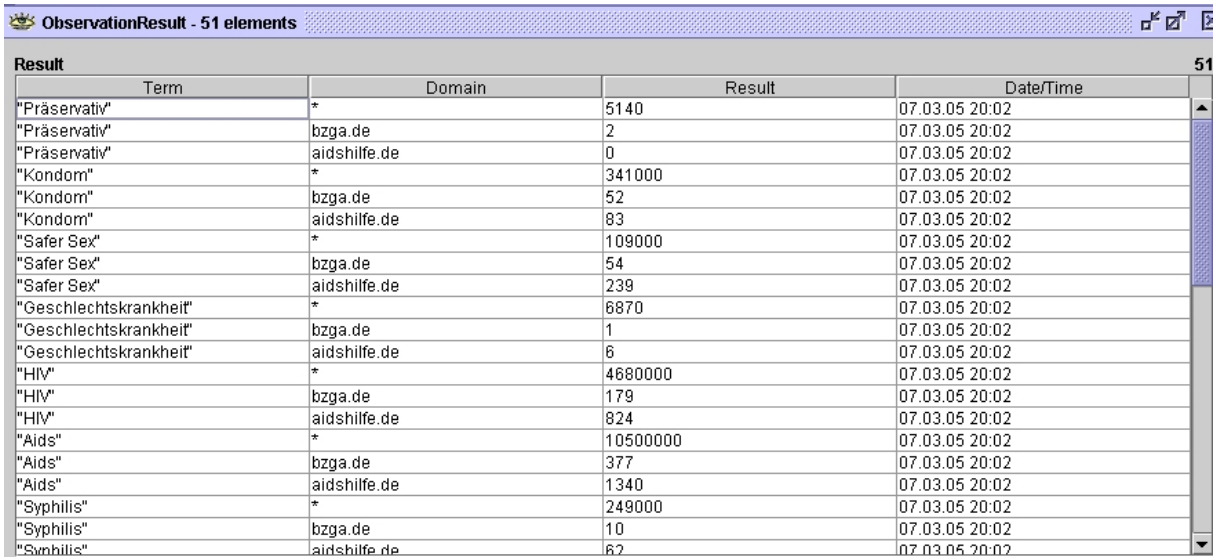


Abbildung 44: ObservationList

Die bisher beschriebenen Schritte werden lediglich ein Mal vom Nutzer durchgeführt. In regelmäßigen Zeitabständen kann anschließend aus der ObservationList ein *ObservationResult* erstellt werden (vgl. Abbildung 45). Dies ist eine Zuordnung von mit Datum versehenen Trefferzahlen zu je einem Begriff in einer Domain. Die dafür notwendigen Daten werden von den Google-WebAPIs geliefert, denen Punkt 6.3 gewidmet ist. ObservationResults können mit anderen ObservationResults zusammengefasst oder nach Kriterien – z.B. einer bestimmten Domain – aufgeteilt werden. Auch ein Exportieren in eine .csv-Datei ist möglich, so dass eine spätere Bearbeitung und Visualisierung durch ein Tabellenkalkulationsprogramm vereinfacht wird.



ObservationResult - 51 elements

Result 51

Term	Domain	Result	Date/Time
"Präservativ"	*	5140	07.03.05 20:02
"Präservativ"	bzga.de	2	07.03.05 20:02
"Präservativ"	aidshilfe.de	0	07.03.05 20:02
"Kondom"	*	341000	07.03.05 20:02
"Kondom"	bzga.de	52	07.03.05 20:02
"Kondom"	aidshilfe.de	83	07.03.05 20:02
"Safer Sex"	*	109000	07.03.05 20:02
"Safer Sex"	bzga.de	54	07.03.05 20:02
"Safer Sex"	aidshilfe.de	239	07.03.05 20:02
"Geschlechtskrankheit"	*	6870	07.03.05 20:02
"Geschlechtskrankheit"	bzga.de	1	07.03.05 20:02
"Geschlechtskrankheit"	aidshilfe.de	6	07.03.05 20:02
"HIV"	*	4680000	07.03.05 20:02
"HIV"	bzga.de	179	07.03.05 20:02
"HIV"	aidshilfe.de	824	07.03.05 20:02
"Aids"	*	10500000	07.03.05 20:02
"Aids"	bzga.de	377	07.03.05 20:02
"Aids"	aidshilfe.de	1340	07.03.05 20:02
"Syphilis"	*	249000	07.03.05 20:02
"Syphilis"	bzga.de	10	07.03.05 20:02
"Syphilis"	aidshilfe.de	62	07.03.05 20:02

Abbildung 45: ObservationResult

Nach dem Sammeln der Daten über einen beliebigen Zeitraum können aus dem ObservationResult passend zu einem TopicGraph mehrere *VisibilityLists* erstellt werden. Für jeden Beobachtungszeitpunkt und jede Domain entsteht hierbei eine VisibilityList (vgl. Abbildung 46). Passend zu einem TopicGraph bedeutet hierbei, dass zu jedem Thema im TopicGraph ein Recallwert vorhanden sein muss. Die Recallwerte des ObservationResult (Integer) werden hierbei in Sichtbarkeitswerte (Float) umgewandelt und je nach Angabe des Benutzers auf einen bestimmten Wert normiert. Neben den Sichtbarkeitsdaten ist in einer VisibilityList Meta-Information gespeichert, beispielsweise über die beobachtete Domain oder den Beobachtungszeitpunkt. Hinzufügen von beliebiger Meta-Information wie ein Benutzername ist möglich. Für zukünftige Erweiterungen ist es denkbar, VisibilityLists auf andere Weise zu impor-

tieren, d.h. nicht aus einem ObservationResult, sondern beispielsweise aus Sichtbarkeitsdaten eines Internetforums.

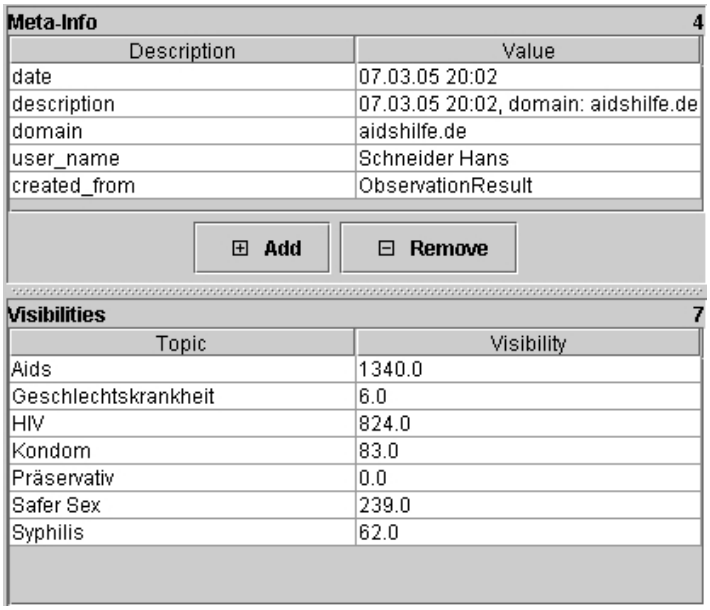


Abbildung 46: VisibilityList

Nach dem Import der Sichtbarkeitsdaten wählt, parametrisiert und startet der Nutzer einen Algorithmus (vgl. Abbildung 47).

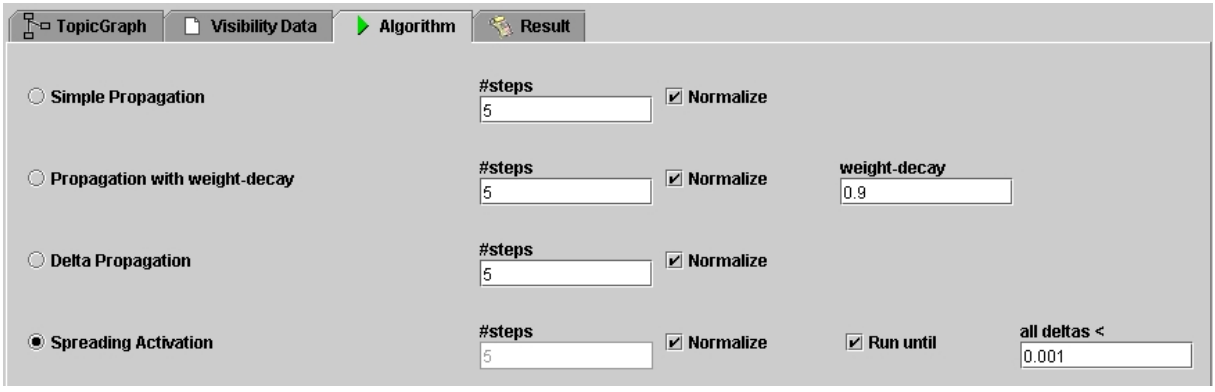


Abbildung 47: Wahl eines Algorithmus'

Das Ergebnis sind wiederum VisibilityLists, hier jedoch mit den abgeleiteten Sichtbarkeiten und zusätzlicher MetaInfo über den gewählten Algorithmus und seine Parameter. Es ist möglich, mehrere verschiedene Algorithmen auf den gleichen VisibilityLists laufen zu lassen, bevor das ganze schließlich in einer *VisibilityTable* gespeichert wird (vgl. Abbildung 48).

Dieser enthält Themen, Sichtbarkeitswerte sowie die Meta-Info aus den VisibilityLists und kann als .csv-Datei zur weiteren Verabeitung gespeichert werden.

VisibilityTable - 63 elements

Visibilities 63

Topic	Visibility	round	steps	algorithm	date	description	domain	created_fr...	normaliza...	weight-de...
Safer Sex	127.1892	4	5	DeltaPropagation	07.03.05 ...	07.03.05 2...	aidshilfe.de	Observati...	2554.0000	-
HIV	713.0602	4	5	DeltaPropagation	07.03.05 ...	07.03.05 2...	aidshilfe.de	Observati...	2554.0000	-
Aids	713.3322	4	5	DeltaPropagation	07.03.05 ...	07.03.05 2...	aidshilfe.de	Observati...	2554.0000	-
Kondom	127.1469	4	5	DeltaPropagation	07.03.05 ...	07.03.05 2...	aidshilfe.de	Observati...	2554.0000	-
Syphilis	32.9466	4	5	DeltaPropagation	07.03.05 ...	07.03.05 2...	aidshilfe.de	Observati...	2554.0000	-
Geschlecht...	713.3282	4	5	DeltaPropagation	07.03.05 ...	07.03.05 2...	aidshilfe.de	Observati...	2554.0000	-
Kondom	248332.5	4	5	Propagation with weight ...	07.03.05 ...	07.03.05 2...	*	Observati...	1589101...	0.899999...
HIV	3516590.5	4	5	Propagation with weight ...	07.03.05 ...	07.03.05 2...	*	Observati...	1589101...	0.899999...
Präservativ	204069.36	4	5	Propagation with weight ...	07.03.05 ...	07.03.05 2...	*	Observati...	1589101...	0.899999...
Geschlecht...	8041022.5	4	5	Propagation with weight ...	07.03.05 ...	07.03.05 2...	*	Observati...	1589101...	0.899999...
Syphilis	9534.606	4	5	Propagation with weight ...	07.03.05 ...	07.03.05 2...	*	Observati...	1589101...	0.899999...

Abbildung 48: VisibilityTable

6.3 Google™-WebAPIs

Zur Ermittlung der Recallwerte werden die Google™-WebAPIs verwendet, die bis heute (10.03.2005) lediglich in der beta-Version verfügbar sind und deren Benutzung auf 1000 Anfragen pro Tag beschränkt ist.

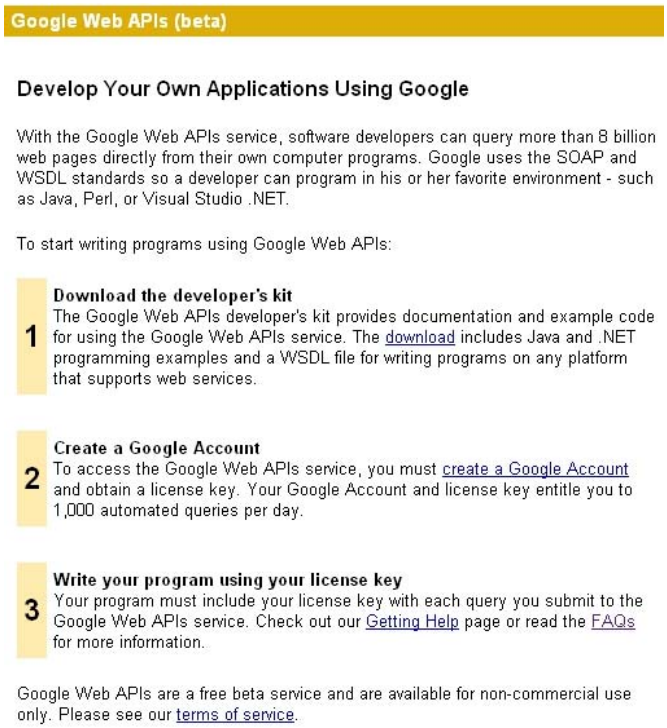


Abbildung 49: Startseite der Google™ WebAPIs (http://www.google.com/apis/, 10.03.2005)

Laut FAQs ist ein Ausbau der Google™-WebAPIs zu einem kostenpflichtigen Service mit mehr als 1000 Anfragen pro Tag und damit die Möglichkeit der kommerziellen Verwendbarkeit im Moment nicht in Planung:

> *"Does Google have any plans to sell Google Web APIs as a service?*
>
> *Not at this time.*
>
> *[…]*
>
> *What if I want to pay Google for the ability to issue more than 1,000 queries per day?*
>
> *Google is only offering the free beta service at this time. If you would like to see Google develop a commercial service, let us know at api-support@google.com."* (http://www.google.com/apis/api_faq.html, 10.03.2005)

Die in Abbildung 49 beschriebenen Schritte zum Einsatz der APIs – Download, Erstellen des Google™-Accounts und Einbau in den eigenen Programmcode – laufen reibungslos und die Verwendung der APIs im Java-Quellcode ist durch mitgelieferte Codebeispiele einfach umzusetzen. Bei Einsatz der APIs ist immer zu beachten, dass lediglich ein geschätzter Wert geliefert wird, denn ab einer bestimmten Größenordnung des Recall-Werts schätzt Google™ die Werte. So ist die in Punkt 2.3.2.1 in Tabelle 1 gelistete Trefferzahl von 10.000.000 für „Aids" sicherlich nicht genau. Doch auch unter Berücksichtigung dieser Schätzungenauigkeiten sind bei Betrachtung der von den APIs gelieferten Daten sporadisch Fehler zu finden, beispielsweise der am 03.03.05 aufgetretene und in Abbildung 50 dargestellte Widerspruch.

ObservationResult - 4 elements

Result 4

Term	Domain	Result	Date/Time
Glatzenrechner	*	515	03.03.05 08:34
Glatzenrechner	alpecin.de	4	03.03.05 08:34
Glatzenrechner	de	1330	03.03.05 08:34
Glatzenrechner	at	107	03.03.05 08:34

Abbildung 50: Ergebnis der Google™-WebAPIs für „Glatzenrechner" in verschiedenen Domains am 03.03.05

Es ist nicht möglich, dass im gesamten Internet (*) nur 515 Treffer für „Glatzenrechner" erzielt wurden, während nach Domaineinschränkung auf aus Deutschland stammende Seiten („de") 1330 Treffer gemeldet werden. Dieser Fehler scheint jedoch nicht an den APIs selbst zu liegen, sondern an den zugrunde liegenden Mechanismen der Suchmaschine Google™,

denn nur zwei Tage nach den Trefferzahlen aus Abbildung 50 liefert die Seite www.google.de folgende geschätzten Trefferzahlen im Rahmen zweier Anfragen (vgl. Abbildung 51):

Abbildung 51: Ergebnis der Suchmaschine Google™ für „Glatzenrechner" in verschiedenen Domains am 05.03.05

Auch hier ergibt sich ein Widerspruch zwischen dem Wert von 998 im gesamten Web und 3180 für Seiten nur aus Deutschland. Es fällt auf, dass der Wert für Deutschland sich kaum verändert hat, während der Wert für das gesamte Web innerhalb von zwei Tagen von 515 auf 998 gesprungen ist. Es liegt also nahe, dass wohl der Recall-Wert für das gesamte Web fehlerhaft ist.

Tabelle 13 zeigt eine andere Stelle, an der widersprüchliche Werte offensichtlich werden: Stark abweichende Werte (nach oben oder unten) von einem Tag auf den nächsten mit sofortiger Rückkehr zum Normalniveau am dritten Tag:

Datum	geschätzte Trefferzahl der Google™-WebAPIs für „Kyoto" in der Domain greenpeace.org
25.01.2005	1780
26.01.2005	2990
27.01.2005	1780

Tabelle 13: Stark abweichende Trefferzahlen der Google-WebAPIs im Verlauf von drei Tagen

Es ist sehr unwahrscheinlich, dass in der Domain greenpeace.org vom 25. Januar bis zum 26. Januar plötzlich 1200 Seiten mehr den Begriff „Kyoto" enthielten und genau die gleiche Anzahl am folgenden Tag wieder verschwunden war. Beim Einsatz der WebAPIs müssen demnach die Daten gefiltert, im Zeitverlauf auffallende Ausreißer identifiziert und beseitigt werden, um die Ergebnisse der Sichtbarkeitsanalyse nicht zu verzerren.

Eine Erklärung der Fehler würde erst durch die Untersuchung des von Google verwendeten Schätzverfahrens bei der Bestimmung der geschätzten Trefferzahlen ermöglicht.

7. EVALUIERUNG

7.1 Nutzen des Einsatzes der Sichtbarkeitsanalyse im Allgemeinen

Zunächst stellt sich die Frage nach dem Nutzen einer Sichtbarkeitsanalyse im Allgemeinen: Ist es tatsächlich so, dass sich Ereignisse aus der Realität im Internet widerspiegeln? Dies kann nur an Hand von Beispielen belegt werden; hierzu sei auf Abschnitt 1 verwiesen (vgl. Abbildung 1, Abbildung 2, Abbildung 4). Die folgende Abbildung 52 zeigt noch ein Beispiel für kalendarische Ereignisse (Weihnachten, Karneval) im Verlauf zwischen Ende Dezember und Ende Februar. Auch hier entsprechen die Kurvenverläufe den Erwartungen, die man auf Grund verschiedener Jahreszeiten an die Sichtbarkeit von Weihnachten und Karneval stellen würde.

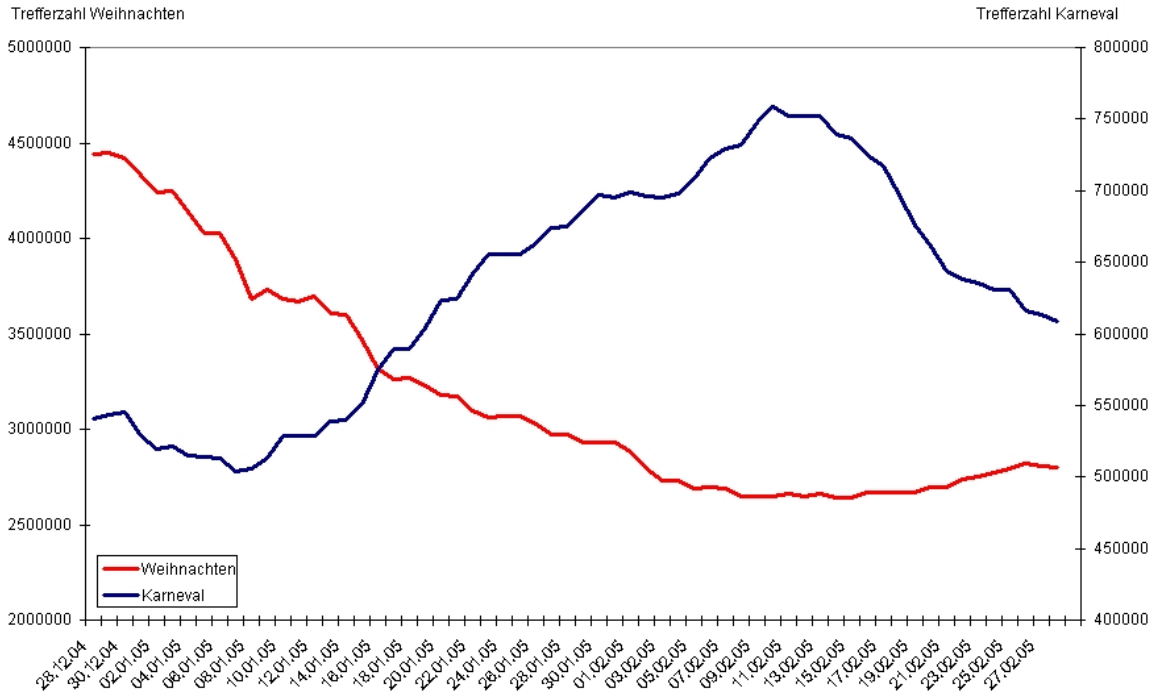


Abbildung 52: Verlauf der Trefferzahlen von „Weihnachten“ und „Karneval“

7.2 Notwendigkeit des Sichtbarkeitsausgleichs in der Realität

Viel entscheidender für diese Arbeit ist die Frage nach dem Nutzen eines Ausgleichs von Sichtbarkeiten: Ist es in realen Beispielen tatsächlich notwendig, einen Sichtbarkeitsausgleich durchzuführen? Bisher wurden hierzu in dieser Arbeit lediglich konstruierte Beispiele mit erfundenen Werten verwendet (vgl. Abschnitt 3). Im Folgenden hierzu noch ein kurzes reales Beispiel zum Thema Klimapolitik.

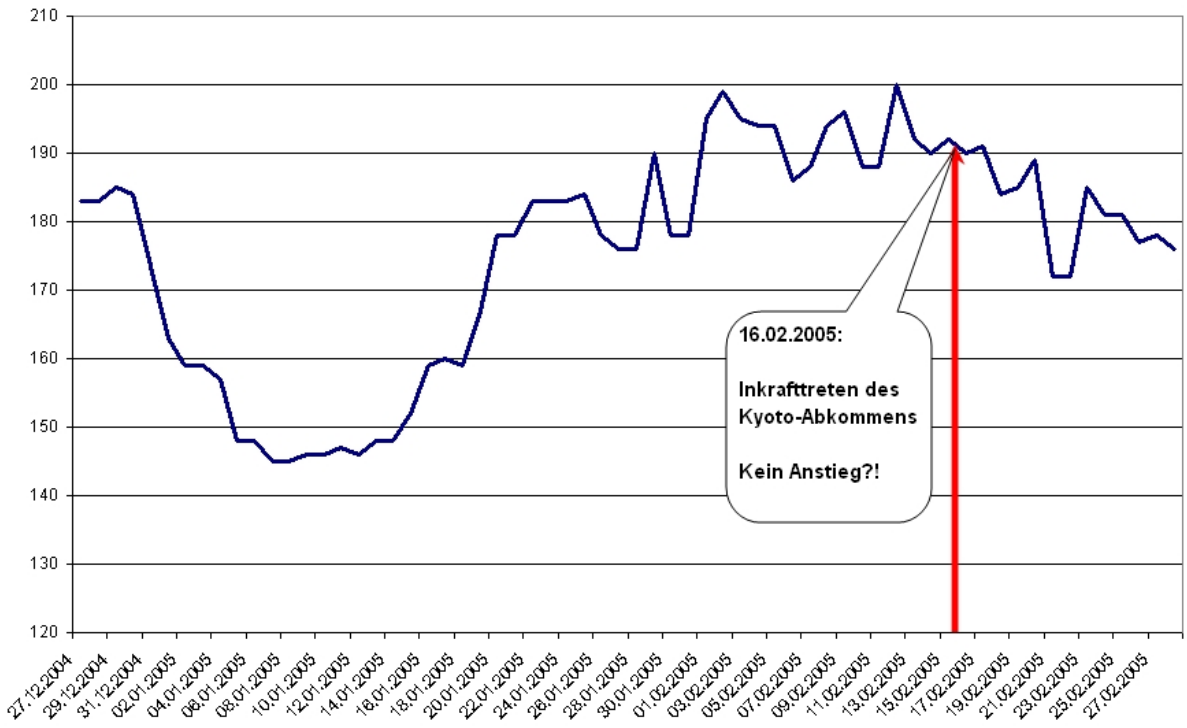


Abbildung 53: Verlauf der Trefferzahlen von „Klimapolitik“ in der Domain greenpeace.org

Abbildung 53 zeigt den Verlauf der Trefferzahlen für „Klimapolitik“ in der Domain www.greenpeace.org von Ende Dezember bis Ende Februar. Bei Inkrafttreten des Kyoto-Abkommens am 16.02.2005 wäre zu erwarten, dass sich dies auf der Webseite einer wichtigen Umweltschutzorganisation wie Greenpeace beobachten ließe. Die Frage: „Ist die Sichtbarkeit des Themas Klimapolitik am 16.2. gestiegen?“ müsste man hier mit „nein“ beantworten und dies als Gegenbeispiel zu der in 7.1 diskutierten allgemeinen Einsetzbarkeit der

Sichtbarkeitsanalyse sehen. Beobachtet man jedoch die Zeitreihe für „Kyoto“[1] in derselben Domain, so ergibt sich folgendes Ergebnis (vgl. Abbildung 54).

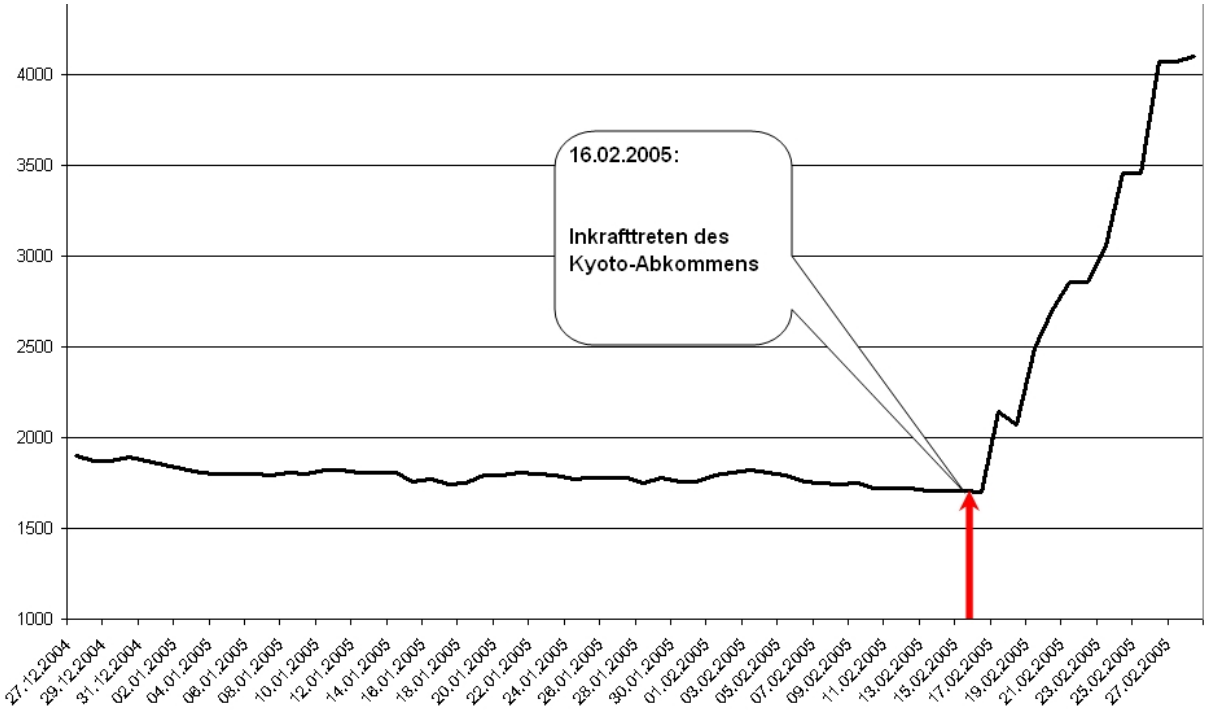


Abbildung 54: Verlauf der Trefferzahlen für „Kyoto“ in der Domain greenpeace.org

Die Werte explodieren am 16. Februar förmlich und verdoppeln sich innerhalb von weniger als zwei Wochen. Dies ist auch einfach zu erklären: Ein Autor, der auf www.greenpeace.org eine aktuelle Meldung zum Kyoto-Protokoll schreibt, wird nur mit einer geringen Wahrscheinlichkeit in dieser Meldung das Wort „Klimapolitik“ verwenden, sicher jedoch das Wort „Kyoto“. Um nun die Frage nach der Sichtbarkeit von „Klimapolitik“ zu beantworten, müssen die beiden Verläufe aus Abbildung 53 und Abbildung 54 zusammengeführt werden, was genau Zielsetzung dieser Arbeit ist. Abbildung 55 zeigt den für dieses Beispiel verwendeten Thesaurus und Abbildung 56 das daraus abgeleitete Themennetzwerk.

[1] Zu beachten ist hier, dass der Suchbegriff „Kyoto“ bei einer Websuche auch viele Seiten einbeziehen wird, die sich mit der Stadt Kyoto und nicht mit Klimapolitik beschäftigen, doch durch die Domaineinschränkung auf www.greenpeace.org kann man davon ausgehen, dass fast alle berücksichtigten Seiten das Kyoto-Abkommen betreffen. Die Suche nach „Kyoto“ anstatt „Kyoto-Abkommen“ soll somit mehrere mögliche Formulierungen gleichzeitig abdecken, wie beispielsweise auch „Kyoto-Protokoll“ oder „Abkommen von Kyoto“.

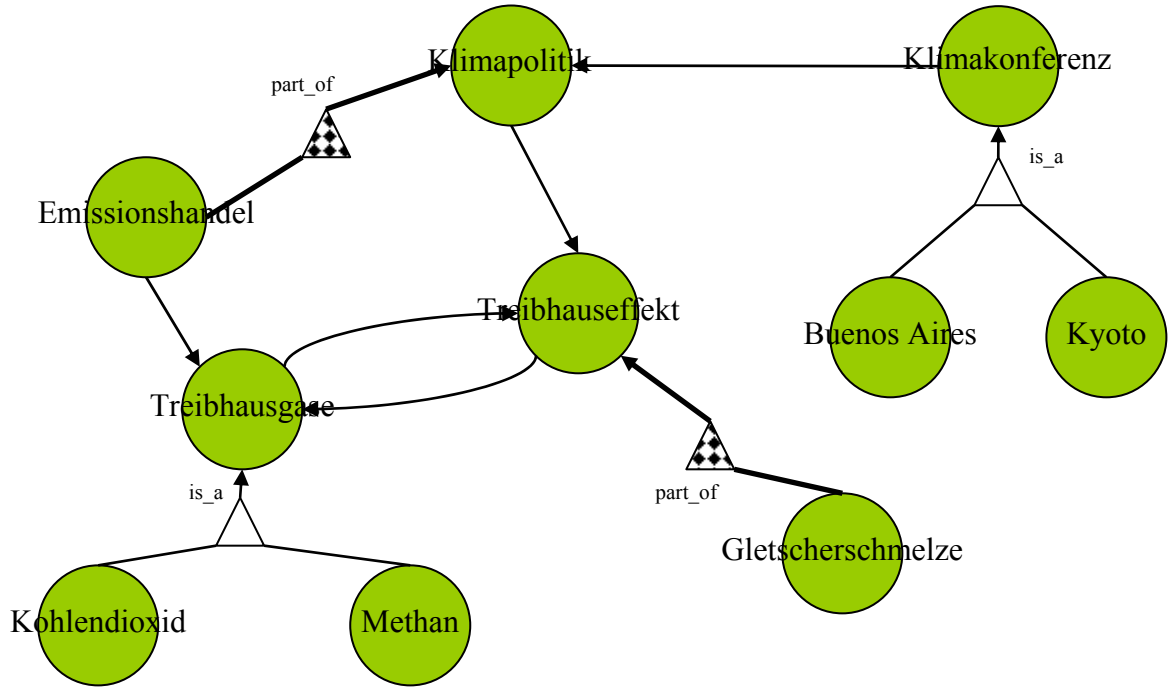


Abbildung 55: Thesaurus zu Klimapolitik

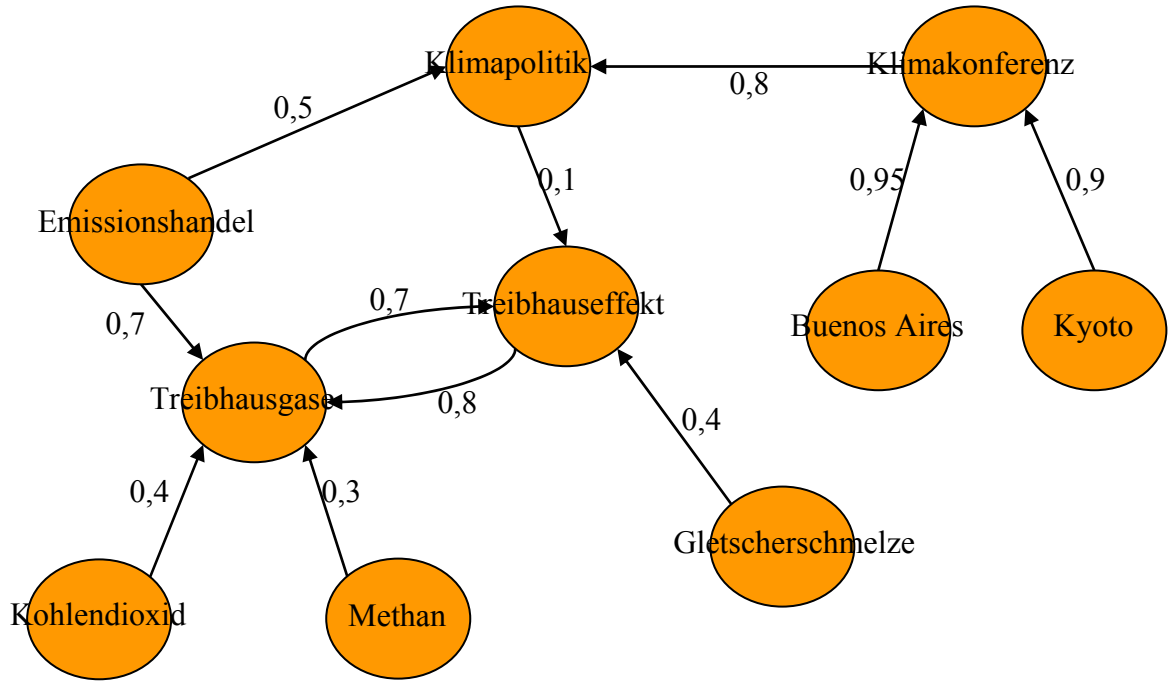


Abbildung 56: Themennetzwerk zu Klimapolitik

Die folgende Abbildung 57 schließlich illustriert den Verlauf der Sichtbarkeiten nach Anwendung der Spreading Activation im Themennetzwerk aus Abbildung 56 im Vergleich zu den reinen Trefferzahlen (= initialen Sichtbarkeiten) aus Abbildung 53 und Abbildung 54.

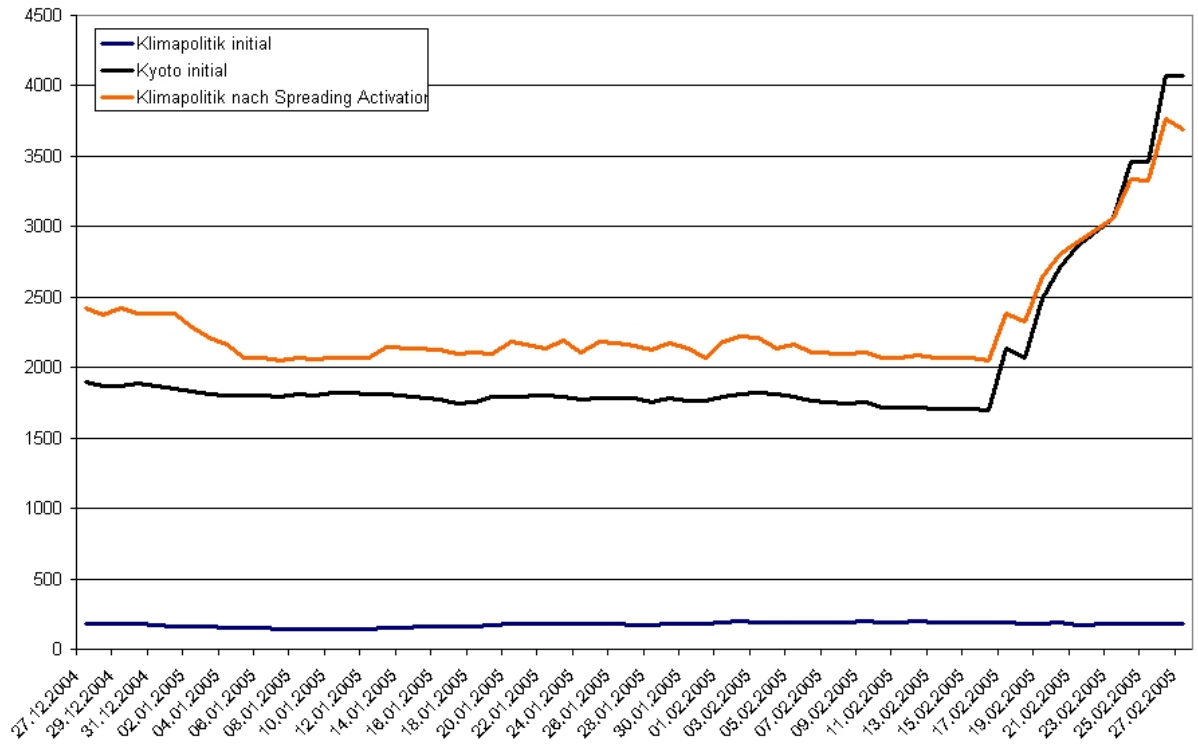


Abbildung 57: Verlauf der initialen und abgleiteten Sichtbarkeit von Klimapolitik

Die Sichtbarkeit von Klimapolitik passt sich durch die Spreading Activation sehr der initialen Kyoto-Kurve an. Das liegt daran, dass im oben stehenden Themennetzwerk die Kantengewichte sehr hoch gewählt wurden und somit die Kyoto-Sichtbarkeit von ca. 2000 fast vollständig bis zur Klimapolitik propagiert wird. Weiterhin fällt auf, dass die initialen Werte im Bereich um die 200 lagen, während sich die abgeleiteten Werte über 2000 befinden. Dieser starke Zuwachs ist im hier gewählten Ansatz erklärbar und auch erwünscht (Sichtbarkeit kann nur zunehmen und Seiten werden entlang einer Kante beim anderen Thema „mitgezählt“). Der Kurvenverlauf der abgeleiteten Sichtbarkeiten beantwortet die Frage „Ist die Sichtbarkeit von Klimapolitik durch das Inkrafttreten des Kyoto-Protokolls gestiegen?“ bei weitem besser als die ursprüngliche Kurve der reinen Trefferzahlen, so dass dies als Beispiel für die Notwendigkeit eines Sichtbarkeitsausgleichs in realen Fällen gesehen werden kann.

7.3 Evaluierung des gewählten Modellierungsansatzes

Eine vollständige Evaluierung des in dieser Arbeit erarbeiteten Ansatzes müsste zunächst den Teil der in Abschnitt 4 vorgestellten Modellierung behandeln. Ist es wirklich sinnvoll, ein

Themennetzwerk aus einem Thesaurus abzuleiten? Insbesondere stellt sich hier die Frage, ob eine direkte Modellierung des Themennetzwerks nicht einfacher wäre und viel Aufwand sparen würde. Diese Frage wurde bereits in Punkt 4.3.2.1 beantwortet: Die gewählte Vorgehensweise bietet dem Anwender eine kognitive Erleichterung. Eine Evaluierung müsste nun diese kognitive Erleichterung durch Versuche mit Testpersonen nachweisen, die eine bestimmte Modellierungsaufgabe entweder mit dem kombinierten Ansatz der Ableitung des Themennetzwerks aus dem Thesaurus oder nur mit einem Themennetzwerk lösen sollen. Das Modellierungsergebnis müsste dann auf Fehlerfreiheit untersucht und eventuell im Verhältnis zur benötigten Modellierungszeit betrachtet werden. Auch eine Untersuchung mit Hilfe eines Laut-Denken-Protokolls zur Analyse des Herangehens der Testpersonen an die Aufgabe ist hierbei möglich.

Eine solche Evaluierung würde den Rahmen dieser Diplomarbeit jedoch sprengen, so dass hierauf an dieser Stelle verzichtet wird.

7.4 Evaluierung der Algorithmen

Die Evaluierung der Algorithmen (DEC), (SA), (SA-A) sowie (SA-B) richtet sich nach den in Punkt 3.2 formulierten und zu Beginn von Abschnitt 5 wiederholten Anforderungen an den Algorithmus zum Sichtbarkeitsausgleich: Anforderungen (1) und (2) werden von allen vorgestellten Algorithmen erfüllt. (SA-A) und (SA-B) berücksichtigen Covisibilitäten und erfüllen somit weiterhin Anforderung (5). Zu testen bleiben demnach noch

(3) Stabilität gegenüber kleinen Veränderungen der Kantengewichte

(4) Stabilität gegenüber kleinen Veränderungen der Anfangswerte

(6) Ergebnisse, die im Rahmen des intuitiv Sinnvollen liegen

und schließlich noch

(7) die in Punkt 5.3.2 gestellte Frage, ob die Spreading Activation mit Covisibilität (SA-A bzw. –B) für reale Covisibilitätswerte entscheidend andere Ergebnisse liefert als die Anwendung von Spreading Activation ohne Covisibilität (SA).

sowie

(8) die Festlegung eines geeigneten Decay-Wertes in (DEC)

Alle hier genannten Untersuchungsziele basieren auf der Frage, wie sich die initialen Sichtbarkeiten durch die Durchführung des jeweiligen Algorithmus' verändert haben. Im nun folgenden Punkt 7.4.1 sollen zunächst zwei Maßzahlen vorgestellt werden, die diese Veränderung zwischen initialen und abgeleiteten Sichtbarkeiten messen.

7.4.1 Messen der Veränderungen im Themennetzwerk

Das Themennetzwerk aus Abbildung 36 diene hier als Beispiel, im Folgenden zwei Mal dargestellt: Vor und nach Durchführung des Algorithmus' (vgl. Abbildung 58).

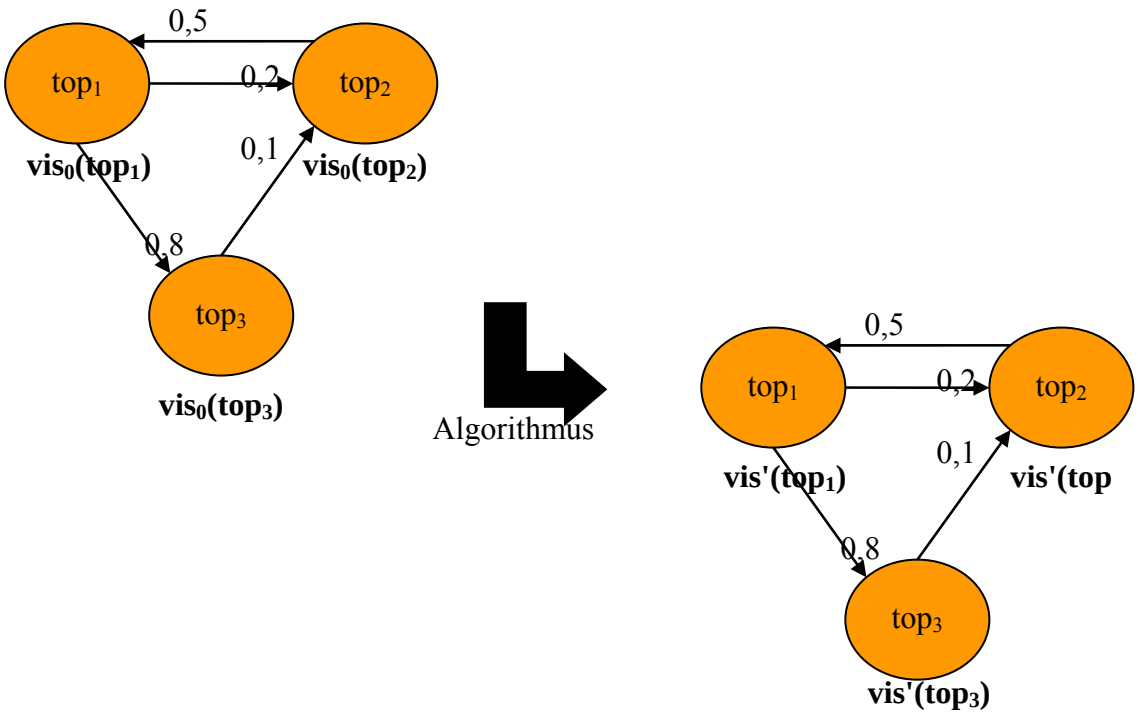


Abbildung 58: Themennetzwerk vor und nach Durchführung eines Algorithmus'

vis$_0$(top) ist eine Abkürzung der bisher in dieser Arbeit verwendeten Bezeichnung visibility$_0$(top) und bezeichnet die initiale Sichtbarkeit eines Themas.

vis'(top) ist entsprechend eine Abkürzung für visibility'(top) und steht für die abgeleitete Sichtbarkeit eines Themas

Δvis(top) = vis'(top) – vis$_0$(top) ist demnach die von einem Thema bei Durchführung des Algorithmus' gewonnene Sichtbarkeit. Δvis(top) ist immer größer gleich 0.

$SUMvis_0 = \sum_{top \in TOP} vis_0(top)$ ist die Summe aller initialen Sichtbarkeiten

$SUMvis' = \sum_{top \in TOP} vis'(top)$ ist die Summe aller abgeleiteten Sichtbarkeiten

$SUM\Delta vis$ = SUMvis' – $SUMvis_0$ ist der Sichtbarkeitsgewinn des gesamten Netzes bei Durchführung des Algorithmus'

$\boldsymbol{prozSUM\Delta vis} = \frac{SUM\Delta vis}{SUMvis_0}$ ist der prozentuale Sichtbarkeitsgewinn des gesamten Netzes bei Durchführung des Algorithmus', abstrahiert also von der Größenordnung der initialen Sichtbarkeiten, da die Algorithmen grundsätzlich auf Sichtbarkeiten um die 100 ebenso wie auf Sichtbarkeitsdaten um die 1000000 angewandt werden können. Dies ist die erste im Folgenden verwendete Maßzahl zum Messen der Veränderungen in einem Themennetzwerk.

$rel\_vis_0(top) = \frac{vis_0(top)}{SUMvis_0}$ ist die relative initiale Sichtbarkeit eines Themas

$rel\_vis'(top) = \frac{vis'(top)}{SUMvis'}$ ist die relative abgeleitete Sichtbarkeit eines Themas

$\Delta rel\_vis(top)$ = rel_vis'(top) – rel_vis_0(top) ist die Veränderung der relativen Sichtbarkeit eines Themas durch Anwendung des Algorithmus'. Diese Zahl wird in Prozentpunkten gemessen und kann sowohl positiv als auch negativ sein. Die Summe aller Δrel_vis(top) in einem Themennetzwerk ist immer 0.

$\sigma^2_{\Delta rel\_vis}$ ist die Varianz über alle Δrel_vis(top) eines Themennetzwerks

$\boldsymbol{\sigma_{\Delta rel\_vis}}$ ist die dazugehörige Standardabweichung und beschreibt, wie stark sich die Sichtbarkeitsverhältnisse im Themennetzwerk durch die Anwendung des Algorithmus' verändert haben.

Beispielsweise ergeben sich für das Themennetzwerk aus Abbildung 58 mit den initialen Sichtbarkeiten aus Abbildung 36 und (SA) folgende in Tabelle 14 dargestellte Werte: Der prozentuale gesamte Sichtbarkeitsgewinn beträgt 127,84%, die Standardabweichung der relativen Sichtbarkeitsänderungen beläuft sich in diesem Beispiel auf 21,25 Prozentpunkte.

$vis_0(top_1)$	100	$vis'(top_1)$	145,93	$\Delta vis(top_1)$	45,93
$vis_0(top_2)$	50	$vis'(top_2)$	91,86	$\Delta vis(top_2)$	41,86
$vis_0(top_3)$	10	$vis'(top_3)$	126,75	$\Delta vis(top_3)$	116,75
$SUMvis_0$	**160**	**SUMvis'**	**364,54**	**SUMΔvis**	**204,54**
				prozSUMΔvis	**127,84%**
$rel\_vis_0(top_1)$	62,50%	$rel\_vis'(top_1)$	40,03%	$\Delta rel\_vis(top_1)$	-22,47%
$rel\_vis_0(top_2)$	31,25%	$rel\_vis'(top_2)$	25,20%	$\Delta rel\_vis(top_2)$	-6,05%
$rel\_vis_0(top_3)$	6,25%	$rel\_vis'(top_3)$	34,77%	$\Delta rel\_vis(top_3)$	28,52%
(SUMME)	**100,00%**	**(SUMME)**	**100,00%**	**(SUMME)**	**0,00%**
				$\sigma_{\Delta rel\_vis}$	**21,25%**

Tabelle 14: Prozentualer Sichtbarkeitsgewinn und Standardabweichung der relativen Sichtbarkeitsänderungen

Variiert man nun je nach Fragestellung einen der Parameter, ändert beispielsweise das Gewicht an einer Kante, eine der initialen Sichtbarkeiten oder den gewählten Algorithmus, so ergeben sich andere Werte für prozSUMΔvis und $\sigma_{\Delta rel\_vis}$. So erhält man für das Themennetzwerk aus Abbildung 58 für ein anderes Kantengewicht von top_1 nach top_2 folgende neuen Werte (vgl. Tabelle 15). Die Untersuchung solcher Veränderungen der Parameter auf die charakteristischen Werte prozSUMΔvis und $\sigma_{\Delta rel\_vis}$ ist die hier zur Evaluierung der Algorithmen gewählte Vorgehensweise.

topicrelationships(top_1, top_2)	**prozSUMΔvis**	**$\sigma_{\Delta rel\_vis}$**
0,2	127,84%	21,25%
0,3	149,23%	20,98%

Tabelle 15: Neue charakteristische Werte für ein Themennetzwerk bei Erhöhung eines Kantengewichts um 0,1

7.4.2 Untersuchte Themennetzwerke

Das Problem von Fragestellungen der Art „Wie verändern sich prozSUMΔvis und $\sigma_{\Delta rel\_vis}$ bei Anwendung eines anderen Algorithmus'?“ ist in erster Linie die Abstraktion vom zugrunde liegenden Netz: Ein Durchlauf arbeitet stets nur auf einer Netzstruktur, doch um eine allgemeine Aussage über einen Algorithmus treffen zu können, müsste davon abstrahiert werden, indem der Algorithmus auf alle möglichen Netzstrukturen angewandt und am Ende beispielsweise ein Mittelwert gebildet wird. Es ist jedoch nicht möglich, alle möglichen Netzstrukturen zu durchlaufen und ebenso ist es unmöglich, ein durchschnittliches Netz zu bilden. Im Bewusstsein dieses Problems bei der Anwendung statistischer Methoden auf Netzwerke soll die Evaluation im Folgenden an Hand ausgewählter Themennetzwerke erfolgen.

(i) Das Themennetzwerk zur Klimapolitik aus Abbildung 56 mit einer typischen Struktur für Themennetzwerke dieser Größe: Mehrere Themen, in die keine Kante hinein-, sondern lediglich eine hinausgeht, da sie aus einer is_a- oder part_of-Beziehung enstanden sind (Kohlendioxid, Methan, Buenos Aires, Kyoto) und ein oder zwei zentrale Themen.

(ii) Das folgende Themennetzwerk (vgl. Abbildung 59) mit vielen Zyklen und einer komplexen Struktur im Kontrast zu (i).

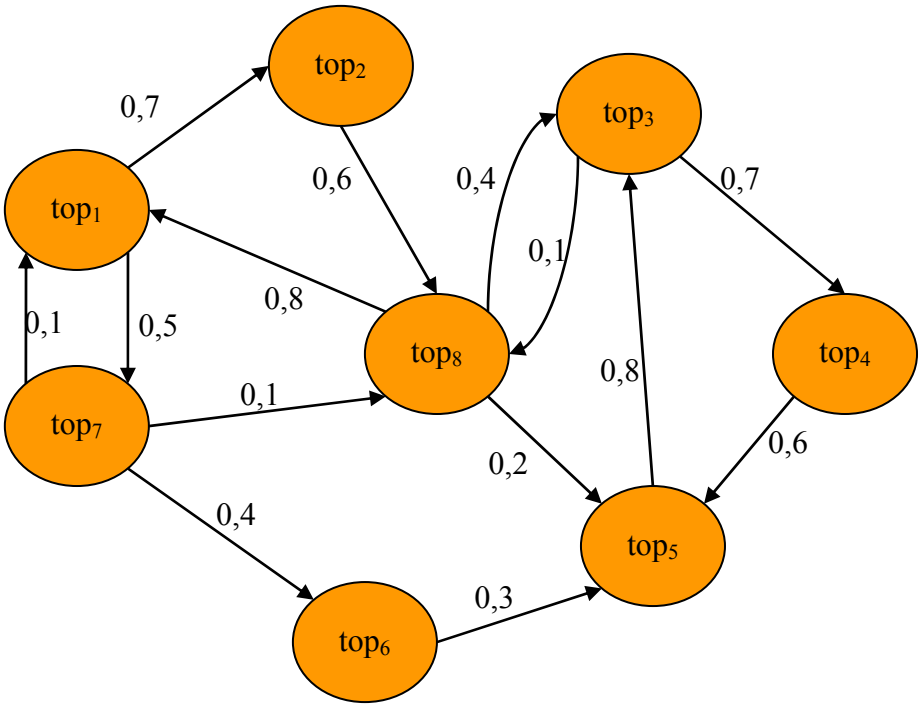


Abbildung 59: Themennetzwerk mit komplexer Struktur

7.4.3 Ergebnisse

7.4.3.1 Verschiedene Decay-Werte im Algorithmus Propagation mit Decay

Die folgenden beiden Diagramme (Abbildung 60 und Abbildung 61) zeigen die Veränderung von prozSUMΔvis für verschiedene Decay-Werte im Algorithmus (DEC) für die Netze (i) und (ii). Hierbei wurden jeweils 50 zufällige Datensätze von initialen Sichtbarkeiten zwischen 0 und 100 erzeugt und jede der 50 Kurven einzeln dargestellt[1]. Der in beiden Diagrammen zu erkennende extreme Anstieg war zu erwarten, denn für decay 1,0 entspricht (DEC) der SimplePropagation, die nicht konvergiert. Decay 0,9 wurde in beiden Diagrammen weggelassen, da der Anstieg von 0,8 zu 0,9 extrem hoch ist. Beim Vergleich der Achsen der beiden Diagramme fällt auf, dass im komplexen Netzwerk (ii) die Gesamtsichtbarkeit viel stärker zunimmt als in (i), was ebenfalls zu erwarten ist: Mehr Kanten führen in (DEC) zu mehr Sichtbarkeitszunahme. Entscheidend ist, dass es keine Ausreißerkurven gibt, d.h. (DEC) zeigt unabhängig von den initialen Sichtbarkeiten das gleiche Verhalten bei Variation des decay-Parameters.

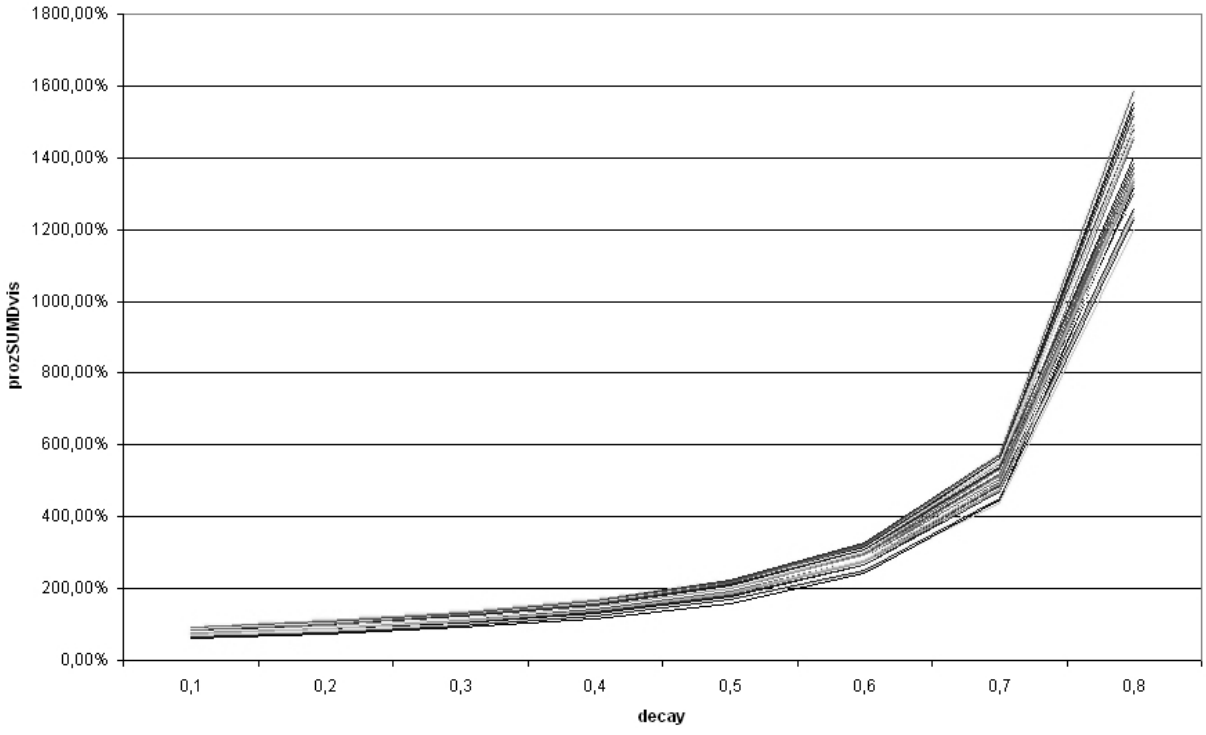


Abbildung 60: prozSUMΔvis im Themennetzwerk (i) mit (DEC) für verschiedene decay-Werte

[1] Jede der Kurven steht also für eine zufällige Belegungdes Themennetzwerks mit initialen Sichtbarkeiten.

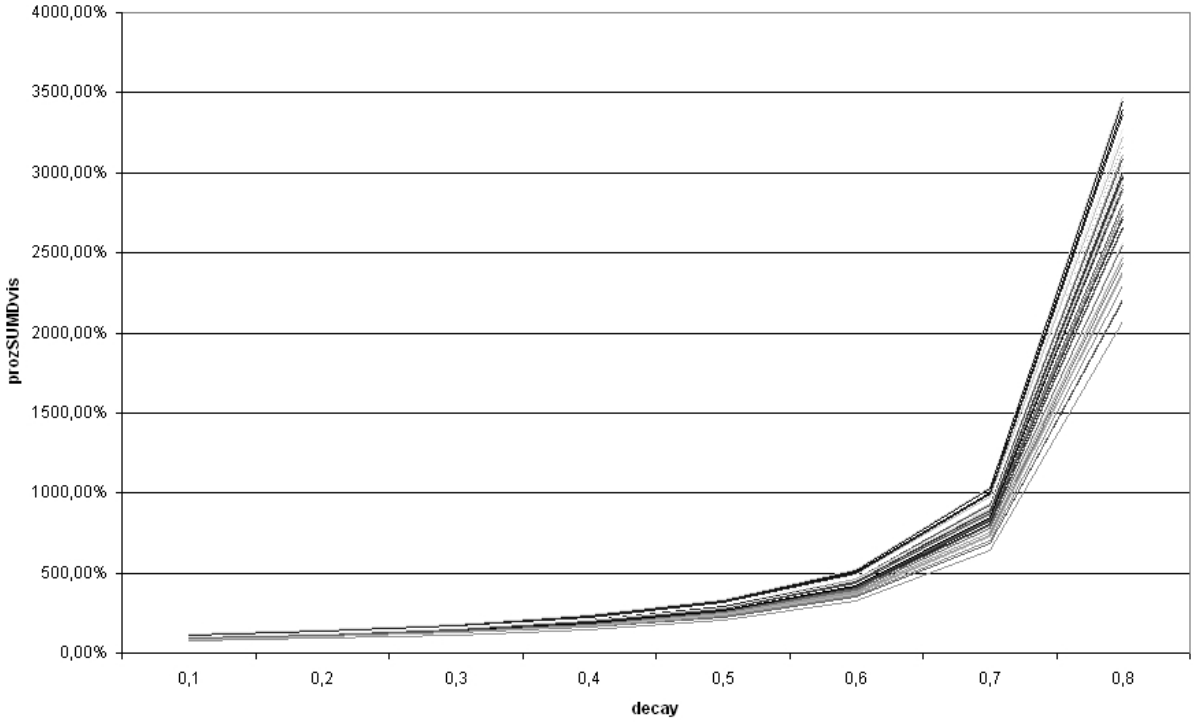


Abbildung 61: prozSUMΔvis im Themennetzwerk (ii) mit (DEC) für verschiedene decay-Werte

Abbildung 62 und Abbildung 63 zeigen $\sigma_{\Delta rel\_vis}$ entsprechend zu den oben stehenden Diagrammen zu prozSUMΔvis.

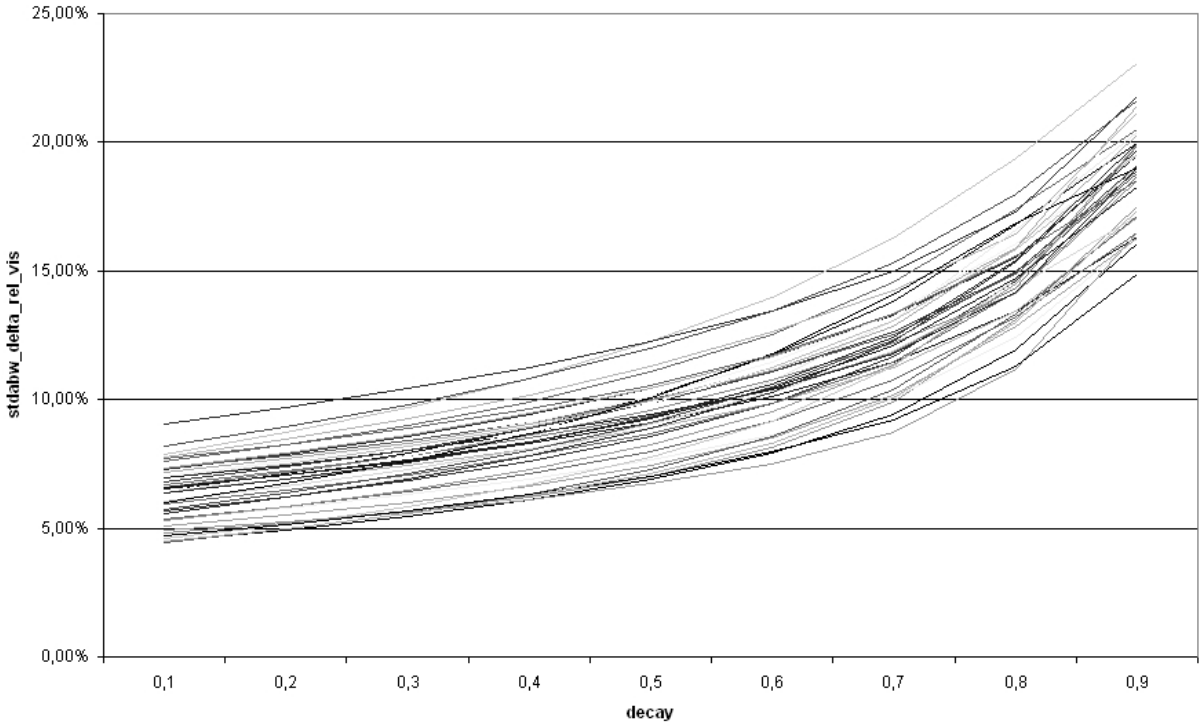


Abbildung 62: $\sigma_{\Delta rel\_vis}$ im Themennetzwerk (i) mit (DEC) für verschiedene decay-Werte

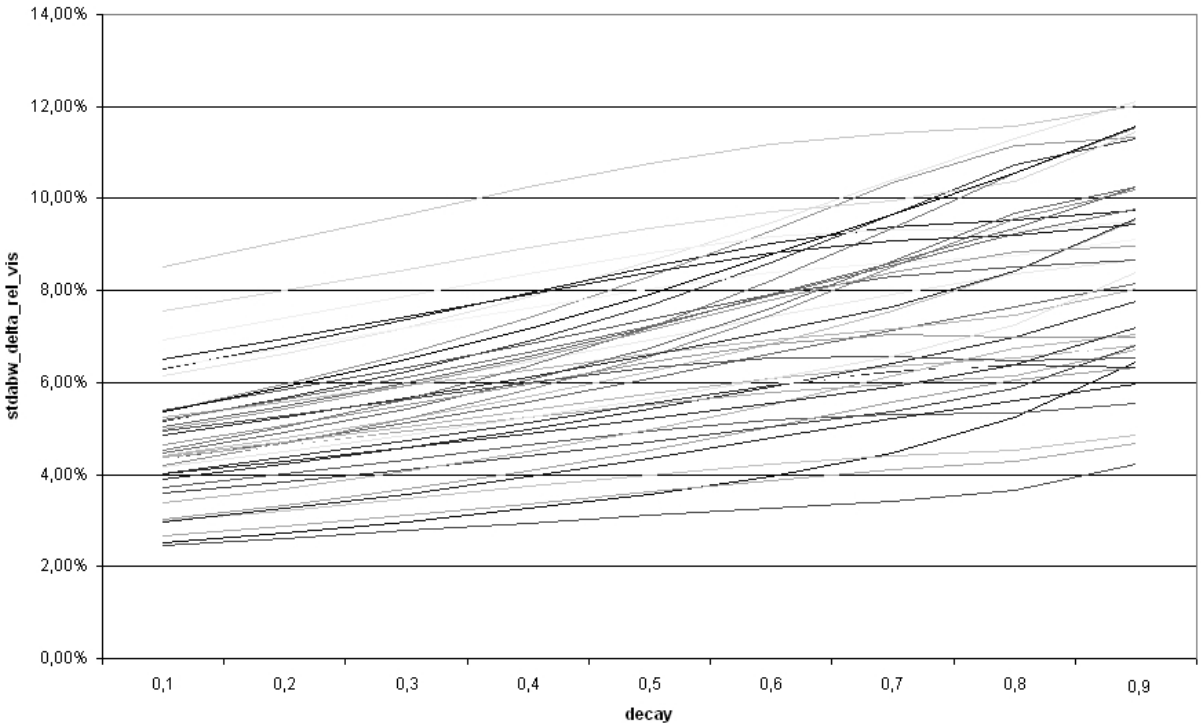


Abbildung 63: $\sigma_{\Delta rel\_vis}$ im Themennetzwerk (ii) mit (DEC) für verschiedene decay-Werte

Hier fällt auf, dass im komplexen Themennetzwerk (ii) die relativen Sichtbarkeiten durch (DEC) allgemein weniger verändert werden als in (i). Das ist damit zu erklären, dass in (ii) jeder Knoten sowohl ein- als auch ausgehende Kanten hat, während es in (i) auch Knoten gibt, die lediglich Sichtbarkeit weitergeben, nicht jedoch welche erhalten, was bei der relativen Betrachtung von Belang ist. Sehr interessant ist auch, dass bei dieser Betrachtung im komplexen Netzwerk (ii) der decay-Parameter eine bei weitem geringere Rolle spielt als in (i): Die Kurven in Abbildung 63 verlaufen sogar fast linear.

Zusammenfassend gesagt hat $\sigma_{\Delta rel\_vis}$ keinen Einfluss auf die Wahl eines geeigneten decay-Parameters. Das bedeutet wiederum, dass bei einem Nutzer, den lediglich relative Sichtbarkeiten interessieren, der gewählte decay nicht relevant ist. Bei einem Nutzer hingegen, der mit den absoluten abgeleiteten Sichtbarkeiten arbeitet, sollte die Wahl des decay-Parameters in einen Bereich fallen, in dem der Anstieg von prozSUMΔvis noch nicht zu hoch ist, d.h. ein decay zwischen 0,3 und 0,6 wäre wohl angemessen. Im Folgenden wird für (DEC) stets decay 0,5 gewählt.

7.4.3.2 Variation von Kantengewichten

a) Variation eines Kantengewichts an einem Blattknoten

Zunächst soll in Netz (i) das Gewicht von einem Blattknoten zum Rest des Netzes, d.h. hier von „Kyoto“ nach „Klimakonferenz“, variiert werden. Die Vorgehensweise entspricht ansonsten der in Punkt 7.4.3.1 gewählten.

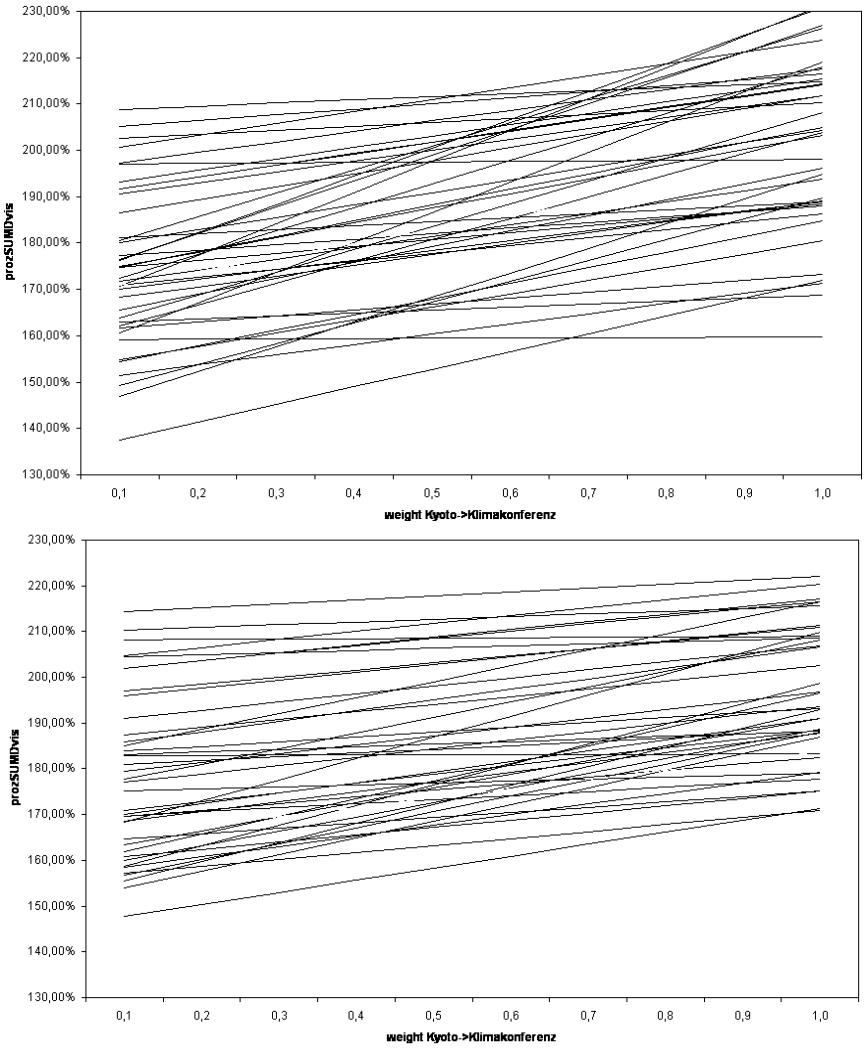


Abbildung 64: prozSUMΔvis im Themennetzwerk (i) für verschiedene Gewichte der Kante von Kyoto nach Klimakonferenz; Algorithmus (DEC) oben und (SA) unten

Abbildung 64 zeigt den Verlauf von prozSUMΔvis im Vergleich zweier Algorithmen: (DEC) und (SA) unterscheiden sich kaum, auch die Größenordnung der Zahlen ist gleich. Die Gesamtsichtbarkeitszunahme entwickelt sich mit zunehmendem Gewicht des Blattknotens linear. Dies ist im Rahmen der Evaluierung positiv zu werten: Eine nur kleine Veränderung des

Kantengewichts zu einem recht unbedeutenden Blattknoten zieht auch keine sprunghaften Veränderungen der Gesamtsichtbarkeitszunahme nach sich.

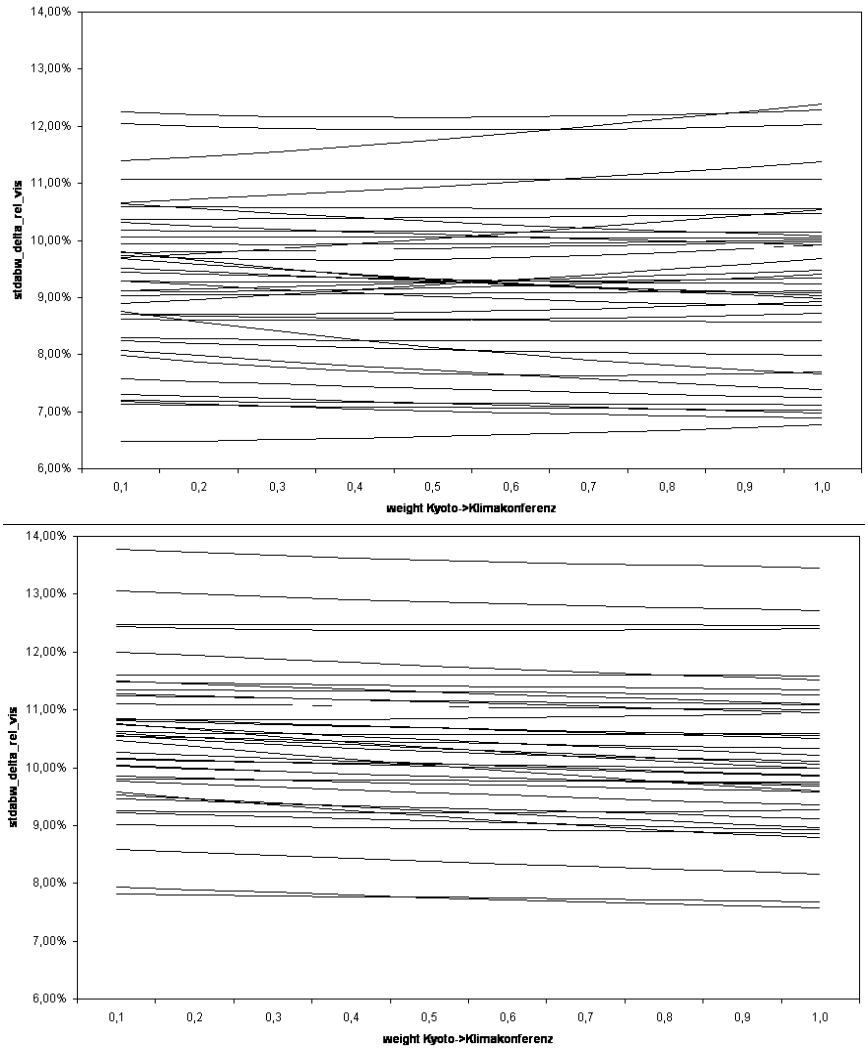


Abbildung 65: $\sigma_{\Delta rel\_vis}$ im Themennetzwerk (i) für verschiedene Gewichte der Kante von Kyoto nach Klimakonferenz; Algorithmus (DEC) oben und (SA) unten

Abbildung 65 zeigt den entsprechenden Verlauf von $\sigma_{\Delta rel\_vis}$. Auch hier befinden sich beide Kurvenscharen in der gleichen Größenordnung. Interessant ist, dass die Kurven fast keine Steigung haben, d.h. dass sich bei Veränderung des Gewichts zum Blattknoten die Sichtbarkeitsverhältnisse im Netz kaum ändern. Dies ist erwünscht, da ein Blattknoten im Netz keine zentrale Rolle einnimmt und daher die Sichtbarkeitsverhältnisse im Gesamtnetz nur am Rande tangieren soll.

b) Variation eines Kantengewichts an einem inneren Knoten

In einem weiteren Test soll nun in Netz (ii) das Gewicht zwischen zwei inneren Knoten, d.h. hier von top_7 nach top_8, variiert werden.

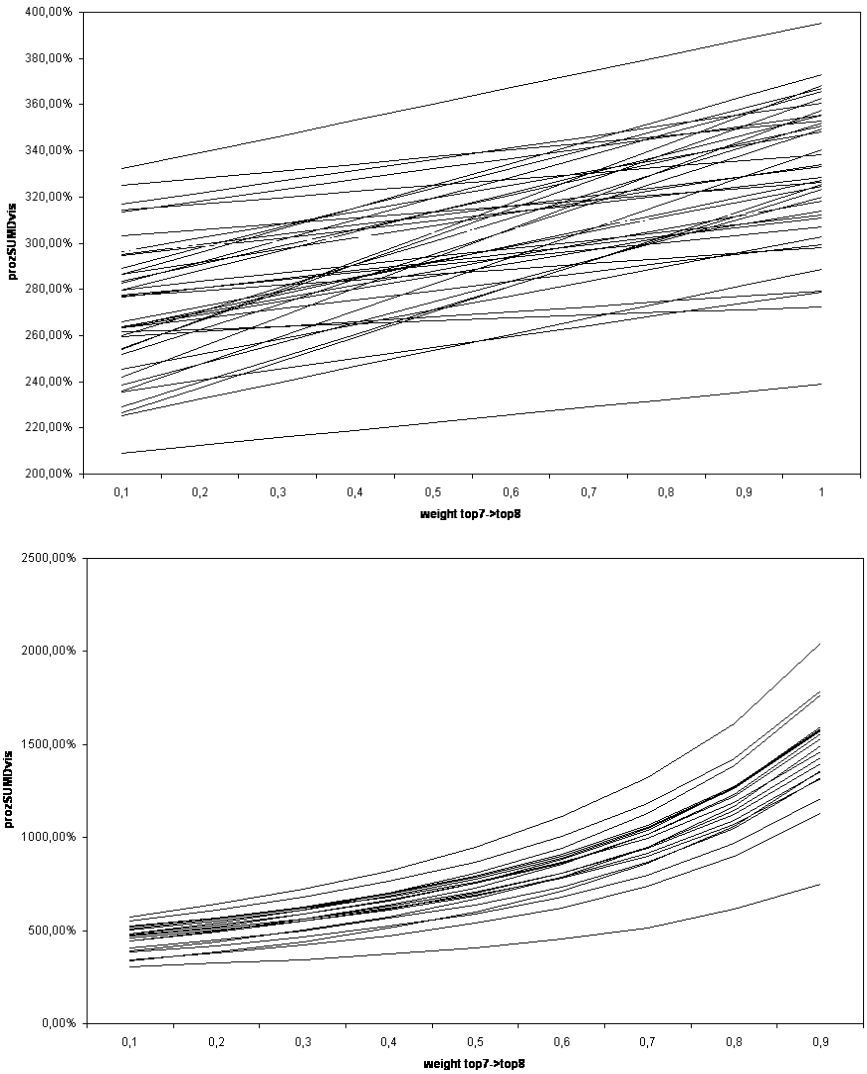


Abbildung 66: prozSUMΔvis im Themennetzwerk (ii) für verschiedene Gewichte der Kante von top_7 nach top_8; Algorithmus (DEC) oben und (SA) unten

Im Gegensatz zu den bisherigen Ergebnissen verändert sich prozSUMΔvis (vgl. Abbildung 66) bei (SA) nicht mehr linear mit steigenden Kantengewichten. Dies liegt am komplexen Netzwerk (ii) mit seinen Zyklen. Auch der Wertebereich ist viel höher als bei (DEC), jedoch bloß beim hier gewählten decay von 0,5 – schon bei einem Decay von 0,7 läge (DEC) bei weitem höher als (SA).

7.4.3.3 Variation von initialen Sichtbarkeiten

a) Variation der initialen Sichtbarkeit eines Blattknotens

In diesem Fall wird in Netz (i) die initiale Sichtbarkeit des Blattknotens „Kyoto“ zwischen 0 und 100 variiert, während alle anderen Sichtbarkeiten zufällig im Bereich 0 bis 100 gewählt werden (vgl. Abbildung 67). Die Kurven für (DEC) und (SA) verlaufen in diesem Fall gleich, so dass hier nur (DEC) dargestellt wird: Eine Veränderung der initialen Sichtbarkeit im Blattknoten Kyoto hat fast keine Auswirkungen.

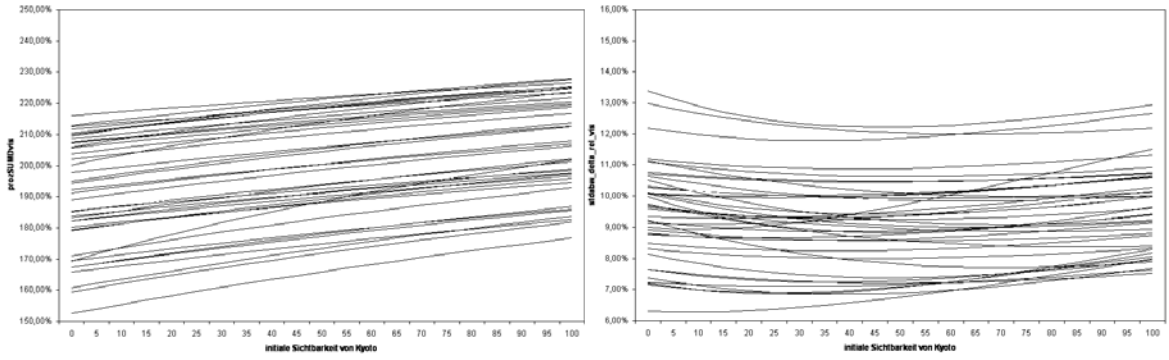


Abbildung 67: prozSUMΔvis (links) und $\sigma_{\Delta rel\_vis}$ (rechts) im Themennetzwerk (i) für initiale Sichtbarkeiten von Kyoto im Intervall [0..100]; Algorithmus (DEC)

Ein zweiter Test (vgl. Abbildung 68) variiert ebenfalls die initiale Sichtbarkeit von „Kyoto“, doch dieses Mal zwischen 0 und 1000, während die anderen initialen Sichtbarkeiten abermals zufällig zwischen 0 und 100 gewählt werden. Dies berücksichtigt den Fall aus Punkt 7.2, wo „Kyoto“ als einziger Knoten im Netz eine sehr hohe initiale Sichtbarkeit (dort über 2000) hatte. Relevant ist in diesem Fall also der Kurvenverlauf ab einer initialen „Kyoto“-Sichtbarkeit von 100. Abbildung 68 zeigt hier für $\sigma_{\Delta rel\_vis}$ einen interessanten Verlauf: Im Bereich 0 bis 100 ist das auch in Abbildung 67 zu sehende leichte Abfallen ablesbar, doch für Werte über 100 steigt $\sigma_{\Delta rel\_vis}$ stark an. Dies ist auch einfach zu erklären: Ein einzelner Blattknoten mit hohem Gewicht gibt diese hohe Sichtbarkeit zu Teilen an das gesamte sonstige Netz ab und erhöht dort sehr stark die initialen Werte, die um die 100 waren. Die Verhältnisse der Sichtbarkeiten zueinander werden also stark verändert, was zu einem hohen $\sigma_{\Delta rel\_vis}$ führt.

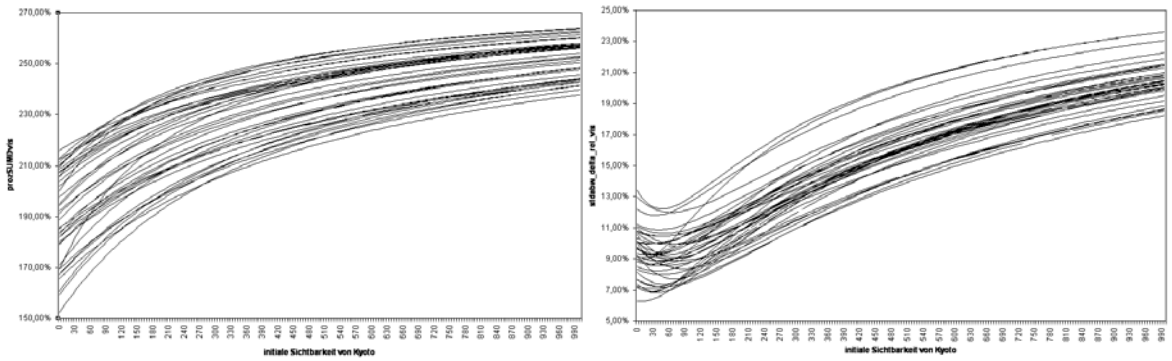


Abbildung 68: prozSUMΔvis (links) und $\sigma_{\Delta rel\_vis}$ (rechts) im Themennetzwerk (i) für initiale Sichtbarkeiten von Kyoto im Intervall [0..1000]; Algorithmus (DEC)

b) Variation der initialen Sichtbarkeit eines zentralen Knotens

Als weiterer Test wird in Netz (ii) die initiale Sichtbarkeit des Knotens top_8 zwischen 0 und 1000 variiert, während die anderen Knoten abermals mit Werten zwischen 0 und 100 zufällig belegt werden (jede der dargestellten Einzelkurven stellt eine dieser Zufallsbelegungen dar). Abermals sind die Kurven für (DEC) und (SA) ähnlich, so dass in Abbildung 69 nur (DEC) dargestellt wird.

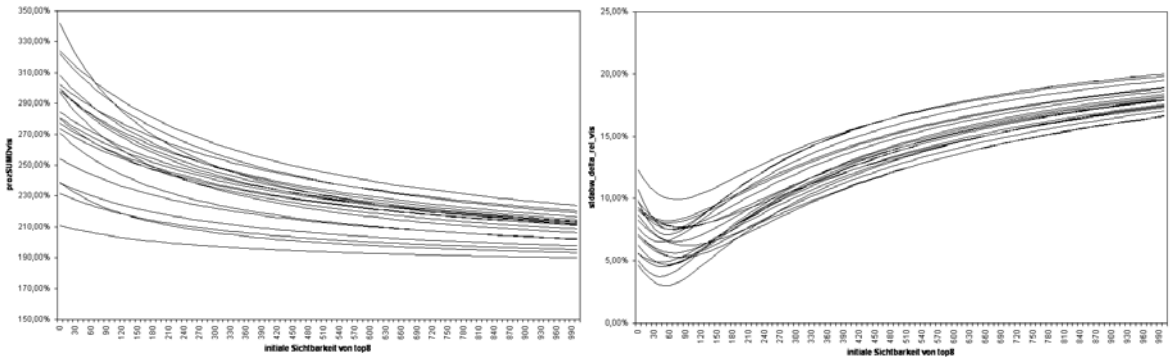


Abbildung 69: prozSUMΔvis (links) und $\sigma_{\Delta rel\_vis}$ (rechts) im Themennetzwerk (ii) für initiale Sichtbarkeiten von top_8 im Intervall [0..1000]; Algorithmus (DEC)

$\sigma_{\Delta rel\_vis}$ verläuft beim zentralen Knoten gleich wie beim Blattknoten. Interessant ist hingegen der Verlauf von prozSUMΔvis, der – anders als bei Abbildung 68 – fallend ist. Ein weiterer Versuch mit Netz (i) und Variation der initialen Sichtbarkeit des inneren Knotens „Klimakonferenz“ zeigt ein fallendes Verhalten gleich zu Abbildung 68. Das bedeutet, dass es nicht an Zyklen oder der allgemein komplizierten Struktur von (ii) liegt, sondern daran, ob der gewählte Knoten Inputkanten hat.

7.4.3.4 Spreading Activation mit oder ohne Covisibilitäten

Die Algorithmen (SA), (SA-A) und (SA-B) sollen hier an einem Beispiel mit realen Covisibilitäten verglichen werden. Ausgangspunkt ist in diesem Fall das Themennetzwerk aus Abbildung 13 in leicht abgeänderter Form (vgl. Abbildung 70): Das Thema „Präservativ“ wurde entfernt.

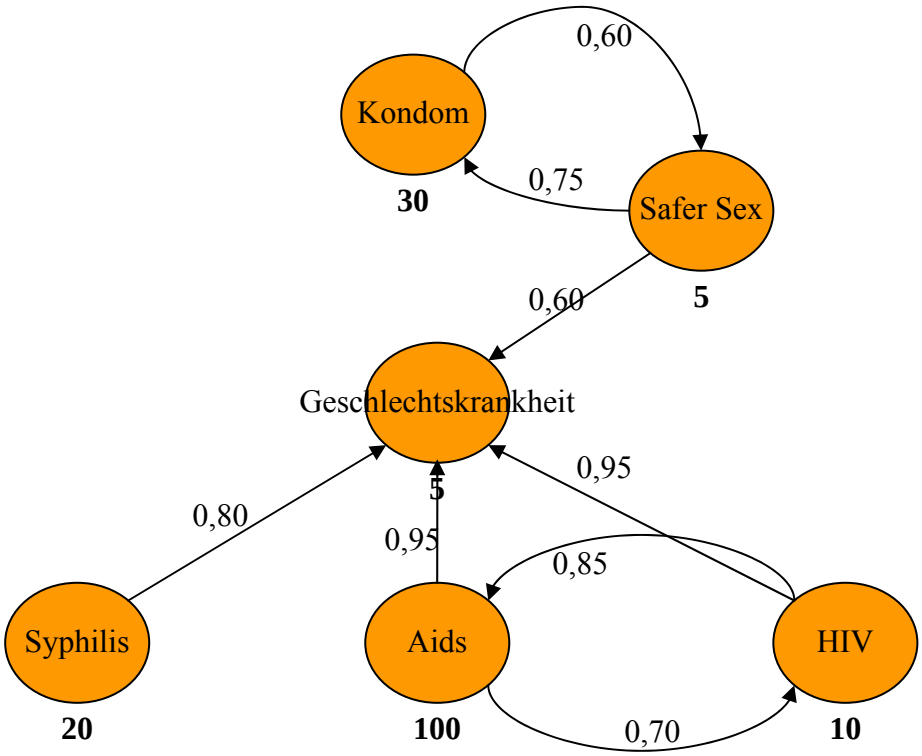


Abbildung 70: Themennetzwerk zum Themengebiet „Safer Sex“ (leicht verändert)

Als initiale Sichtbarkeiten sowie Covisibilitäten werden reale Trefferzahlen vom 06.03.05 verwendet (vgl. Tabelle 16, Tabelle 17, Tabelle 18).

No	Thema	initiale Sichtbarkeit
1	Kondom	342000
2	Safer Sex	108000
3	Geschlechtskrankheit	6850
4	HIV	7110000
5	Aids	10500000
6	Syphilis	312000

Tabelle 16: Initiale Sichtbarkeiten zum Themengebiet „Safer Sex“ vom 06.03.05

	1	2	3	4	5	6
1	*342000*	4770	496	X	X	X
2		*108000*	333	X	X	X
3			*6850*	1330	1670	1440
4				*7110000*	2620000	X
5					*10500000*	X
6						*312000*

Tabelle 17: Co-Recallwerte zum Themengebiet „Safer Sex“ vom 06.03.05

Die Co-Recallwerte in Tabelle 17 sind wie folgt zu interpretieren: Felder, in denen ein X steht, werden nicht benötigt, da kein Pfad zwischen diesen Themen besteht und daher in den Algorithmen die Covisibilität zwischen diesen Themen nicht relevant ist. Der Co-Recallwert eines Themas zu sich selbst ist trivialerweise der Recallwert des Themas, d.h. die initiale Sichtbarkeit aus Tabelle 16. Tabelle 18 listet die aus diesen Daten berechnete covisibility$_1$(top$_1$, top$_2$), wobei top$_1$ in den Zeilen und top$_2$ in den Spalten zu finden ist.

	1	2	3	4	5	6
1	*1*	0,013947	0,00145	X	X	X
2	0,044167	*1*	0,003083	X	X	X
3	0,072409	0,048613	*1*	0,194161	0,243796	0,210219
4	X	X	0,000187	*1*	0,368495	X
5	X	X	0,000159	0,249524	*1*	X
6	X	X	0,004615	X	X	*1*

Tabelle 18: covisibility$_1$ zum Themengebiet „Safer Sex“ vom 06.03.05

Nach Durchführung der verschiedenen Algorithmen ergibt sich das folgende Bild für die relativen Sichtbarkeiten (vgl. Abbildung 71): Jeder der Algorithmen hat die relativen Sichtbarkeiten im Vergleich zu den initialen Sichtbarkeiten stark verändert, doch ein Unterschied zwischen den Werten von (SA), (SA-A) und (SA-B) im Bezug auf die relativen Sichtbarkeiten ist kaum zu erkennen. Das zeigt sich auch an den fast gleichen $\sigma_{\Delta rel\_vis}$-Werten für die drei Algorithmen:

(SA): $\sigma_{\Delta rel\_vis} = 24{,}08719\%$

(SA-A): $\sigma_{\Delta rel\_vis} = 24{,}10531\%$

(SA-B): $\sigma_{\Delta rel\_vis} = 24{,}13108\%$

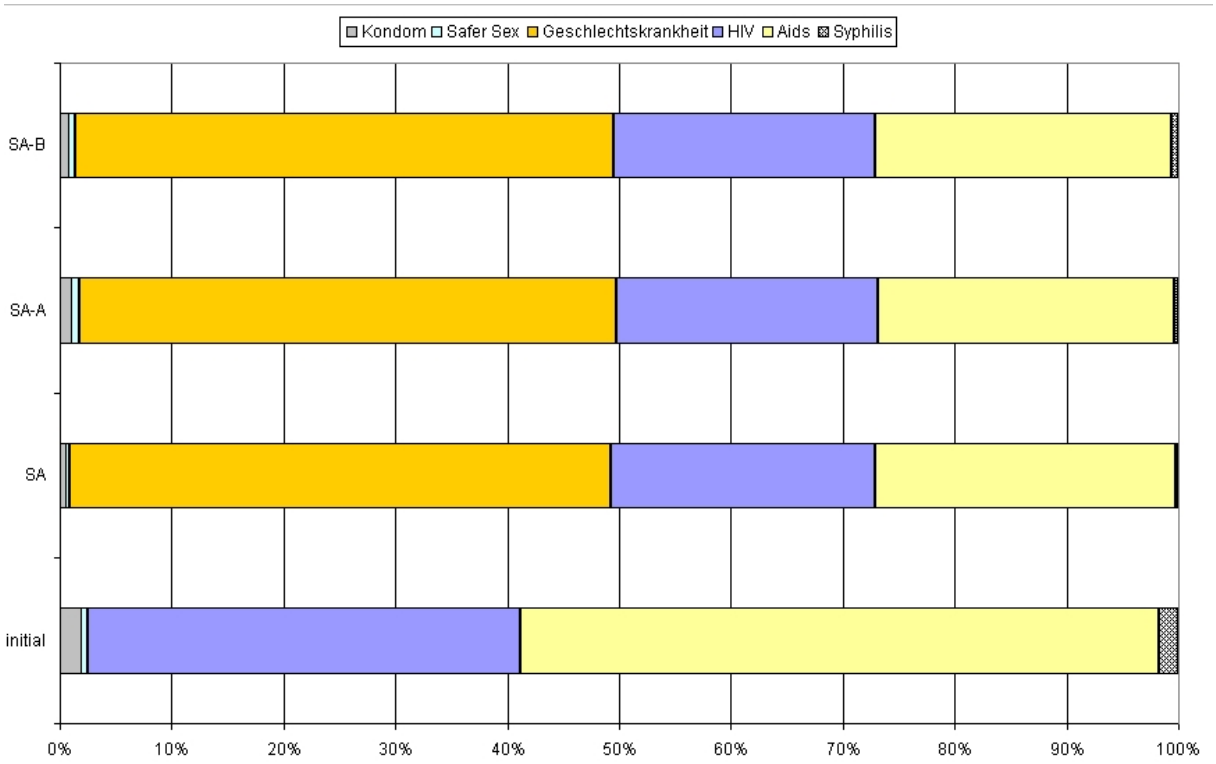


Abbildung 71: Relative Sichtbarkeiten: initial, (SA), (SA-A) und (SA-B)

Der entscheidende Unterschied ist an den Werten für prozSUMΔvis zu erkennen:

(SA): prozSUMΔvis = 724,45%

(SA-A): prozSUMΔvis = 309,83%

(SA-B): prozSUMΔvis = 193,91%

Die Verwendung von Covisibilitäten hat also einen stark bremsenden Einfluss auf die absolute Höhe der Werte. Dies war zu erwarten, da bei jedem Propagierungsschritt nicht nur mit dem Kantengewicht, sondern auch mit (1-$covis_1$) multipliziert wird. Dass (SA-B) die Gesamtsichtbarkeitszunahme noch mehr hemmt als (SA-A) liegt an der in 5.3.2 erklärten Besonderheit von (SA-B): Ein ewiges Durchlaufen von Zyklen ist wegen $covis_1(top_A, top_A)$ nicht möglich – spätestens, wenn die von top_A ausgehende Sichtbarkeit einmal im Kreis herumgereicht wurde und wieder bei top_A ankommt, wird dieser Teil der Sichtbarkeit nicht mehr weiter propagiert.

7.4.3.5 Zusammenfassung

Im Bewusstsein, dass eine umfassende Evaluierung der Algorithmen nur durch Testen noch anderer verschiedener Netze möglich ist und hier sicher nicht alle Aspekte beleuchtet werden konnten, lassen sich die Ergebnisse aus Punkt 7.4.3 folgendermaßen zusammenfassen:

- Der Parameter decay im Algorithmus (DEC) hat einen überaus großen Einfluss auf die absolute Höhe des Ergebnisses. Eine kleine Veränderung dieses Parameters führt in dem Bereich ab ca. decay=0,65 zu sehr großen Sichtbarkeitsveränderungen. Daher sollte ein decay im Intervall [0,3..0,6] gewählt werden.
- Eine kleine Veränderung eines Kantengewichts an einem Blattknoten führt zu kleinen Veränderungen der Gesamtsichtbarkeitszunahme sowie der Sicherbarkeitsverhältnisse.
- Eine kleine Veränderung eines Kantengewichts an einem inneren Knoten kann v.a. im Algorithmus (SA) (aber auch in (DEC) mit bestimmten decay-Werten) zu einer großen Veränderung der Gesamtsichtbarkeitszunahme führen. Auf die Höhe der Kantengewichte bei inneren Knoten (insb. bei Zyklen) ist demnach bei der Modellierung besondere Rücksicht zu nehmen. Zu hohe Gewichte sind eher zu vermeiden.
- Eine kleine Veränderung der initialen Sichtbarkeit eines Knotens führt zu kleinen Veränderungen der Gesamtsichtbarkeitszunahme sowie der Sichtbarkeitsverhältnisse.
- Knoten, die in ihrer initialen Sichtbarkeit weit über dem Durchschnitt der anderen Knoten liegen, beeinflussen die relativen Sichtbarkeitsverhältnisse stark.
- Reale Covisibilitäten können hoch genug sein (vgl. HIV, Aids), um einen gravierenden Einfluss zu haben, d.h. (SA-A) oder (SA-B) unterscheiden sich in ihren Ergebnissen für reale Covisibilitäten stark vom einfachen (SA).
- Die Covisibilitäten haben bremsende Wirkung und können so ein Explodieren der Werte in zyklischen Strukturen verhindern.
- (SA-B) erfüllt alle Anforderungen aus Punkt 3.2, da bei ihm (neben der Erfüllung der anderen Anforderungen) die Ergebnisse durch die stark bremsende Wirkung am ehesten im Bereich des intuitiv Sinnvollen liegen.

8. ERGEBNISSE UND AUSBLICK

Der erste im Rahmen dieser Arbeit zu erwähnende interessante Aspekt sind die realen Sichtbarkeitsverläufe der Fallbeispiele: Es lässt sich tatsächlich ein Zusammenhang zwischen aktuellen Ereignissen, gesellschaftlichen Strömungen oder Werbekampagnen und der Sichtbarkeit eines Themas im Internet feststellen – und dies schon bei einem derart einfach definierten Maß wie der Trefferzahl einer Suchmaschine. Der Einsatz der Sichtbarkeitsanalyse birgt also tatsächlich Potential zu Forschungszwecken sowie solches wirtschaftlicher Art.

Der Schwerpunkt dieser Arbeit lag im Sichtbarkeitsausgleich, d.h. in der nachträglichen Korrektur von Sichtbarkeiten durch Berücksichtigung von semantischen Beziehungen zwischen Themen. Die Notwendigkeit eines solchen Ausgleichs wurde in 7.2 an einem weiteren Fallbeispiel belegt. Ebenfalls wurde an diesem Fallbeispiel gezeigt, wie der in dieser Arbeit gewählte Ansatz der Spreading Activation auf einem Themennetzwerk den erwünschten Sichtbarkeitsausgleich ermöglicht.

Im Vergleich verschiedener Algorithmen sind Propagation mit Decay und die Spreading Activation als fast gleichwertig zu sehen, jedoch nur bei geeignetem Parameter decay für ersteren Algorithmus. Auf Grund dieser Einschränkung und der in 5.2.4.3 dargelegten Überlegungen zur intuitiven Begründung für Spreading Activation an Hand der Vorstellung des „Mitzählens“ von Seiten bei semantischen verwandten Themen wird Spreading Activation präferiert. Das grundsätzliche Problem aller vorgestellten Algorithmen ist die immense Zunahme der Sichtbarkeiten, die durch den Hauptbeitrag dieser Arbeit – der Abwandlung der Spreading Activation zur Berücksichtigung der Covisibilitäten – insbesondere in jenen Fällen gelöst wird, in denen hohe Covisibilität zwischen Themen besteht. Doch auch bei niedrigen Covisibilitäten führt der vorgestellte Algorithmus (SA-B) durch Vermeiden zyklischer Sichtbarkeitsweitergaben zu annehmbaren Werten.

Für den zukünftigen Einsatz der Sichtbarkeitsanalyse ist auf der einen Seite der fachliche Aspekt des jeweiligen Einsatzgebietes interessant. Hier bieten sich noch ungeahnte Möglichkeiten zu Studien in verschiedenen Bereichen wie der Kommunikationswissenschaft oder beispielsweise Untersuchungen zur Eignung der Sichtbarkeitsanalyse im Marketing. Insbesondere bietet sich ein Vergleich des Sichtbarkeitsverlaufs von Themen im Internet mit dem in Printmedien oder dem Fernsehen an. Ein für viele Anwendungsfälle wichtiger Aspekt ist weiterhin die Frage, auf welchen Webseiten das Thema sichtbar ist. So wäre es sicherlich nicht im Sinne einer Analyse der Auswirkungen von Marketingmaßnahmen im Alpecin®-Beispiel, falls sämtliche Seiten mit „Coffein-Complex“ vom Hersteller selbst stammten.

Auch aus Sicht der Informatik bleiben ungelöste Fragen: So ist die in 2.1 getroffene Einschränkung der Sicht eines Themas als Zeichenkette sicherlich stark vereinfachend und sollte in Richtung einer weiteren Zuordnung von Eigenschaften verändert werden; zumindest Synonyme sollten darstellbar sein. Auch in diese Richtung der Zuordnung von Eigenschaften tendiert die in 4.2.2 bei den Topic Maps bereits vorgeschlagene Möglichkeit der Modellierung beliebiger Beziehungen mit Spezifizierung der Eigenschaften des jeweiligen Beziehungstyps (z.B. Transitivität). Eine Erweiterung der Definition der Covisibilität auf n gemeinsam sichtbare Themen sowie der Einsatz dieser erweiterten Covisibilitäten im jeweiligen Algorithmus sind denkbar. Aus algorithmischer Sicht wäre es außerdem interessant, direkt auf einem Thesaurus zu arbeiten und hierbei mit mehreren Beziehungstypen umzugehen – fünf im hier vorgestellten Thesaurus, beliebig viele bei Kombination mit oben erwähnter beliebiger Modellierung der Beziehungstypen. Auch die in Punkt 4.3.2.2 angedeutete Unterstützung der Gewichtsfestlegung durch Covisibilitäten – sprich: Die heuristische Ermittlung der semantischen Nähe zwischen Themen auf Grund ihres gemeinsamen Auftretens in Webdokumenten – könnte Gegenstand weiterer Untersuchungen sein.

Schließlich bleiben noch erweiterte Möglichkeiten der Ermittlung von initialen Sichtbarkeitsdaten. Dies wurde hier durch die primitive Methode des reinen Zählens von Webseiten durchgeführt, da im Rahmen eines Sichtbarkeitsausgleichs initiale Sichtbarkeitsdaten lediglich als Input dienen. Eine interessante Variante ist hier der Import von Sichtbarkeitsdaten aus anderen Internetprozessen wie Internetforen, aber auch die in 2.2.1.1 in Form von Anmerkungen erwähnten genaueren Möglichkeiten zur Gestaltung einer Relevanzfunktion sollten in Be-

tracht gezogen werden. Beispielsweise wäre hier auch die Berücksichtigung der Struktur einer Webseite denkbar, d.h. ein Messen des Abstands einer Unterseite von der Startseite. Im Gegensatz zum einfachen Zählen der Seiten könnte eine Gewichtung der Trefferzahlen mit dem jeweiligen PageRank der Seite zu einem genaueren Sichtbarkeitsmaß führen.

Insgesamt konnte im Rahmen dieser Diplomarbeit nur ein Teilaspekt der variantenreichen Möglichkeiten der Sichtbarkeitsanalyse von Themen im Internet angegangen werden. Eine Verstärkung der Forschungstätigkeiten auf diesem Gebiet bleibt wünschenswert.

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

LITERATURVERZEICHNIS

[BaezaRib04]
Baeza-Yates, R., Ribeiro-N, B. 2004. Modern Information Retrieval, *ACM Press [u.a.], New York [u.a.]*

[BerHenLas01]
Berners-Lee, T., Hendler, J., Lassila, O. 2001. The Semantic Web, *Scientific American, May 17, 2001*

[BrinPage98]
Brin, S., Page, L. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer Networks and ISDN Systems*

[Broder&al00]
Broder, A. et al. 2000. Graph structure in the web, *Computer Networks: The International Journal of Computer and Telecommunications Networking, Volume 33, Issue 1-6 (June 2000)*

[CegCobCua03]
Ceglowski, M., Coburn, A., Cuadrado, J. 2003. Semantic Search of Unstructured Data using Contextual Network Graphs, *National Institute for Technology and Liberal Education*

[CollLoft75]
Collins, A.M., Loftus, E.F. 1975. A Spreading-Activation Theory of Semantic Processing, *Psychological Review 1975, Vol. 82, No. 6, S. 407-428*

[Connolly&al01]
Connolly, D., van Harmelen, F., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., Stein, L.A. 2001. DAML+OIL (March 2001) reference description, *W3C Note, 18 December 2001* (http://www.w3.org/TR/2001/NOTE-daml+oil-reference-20011218, zuletzt besucht am 23.02.2005)

[Cruse95]
Cruse, D. A. 1995. Lexical semantics, *Cambridge University Press, Cambridge [u.a.]*

[Diestel00]
Diestel, R. 2000. Graphentheorie, *Springer, Berlin [u.a.]*

[DIN1463]
Deutsches Institut für Normung e.V.. 1987. DIN 1463 Teil 1, Erstellung und Weiterentwicklung von Thesauri. Einsprachige Thesauri

[Falkinger03]
Falkinger, J. 2003. Attention Economies, *CESIFO WORKING PAPER NO. 1079, ifo Institut für Wirtschaftsforschung, München*

[Fellbaum98]
Fellbaum, C. (ed.) 1998. WordNet, an electronic lexical database, *MIT Press, Cambridge, Mass [u.a.]*

[Fensel&al01]
Fensel, D., van Harmelen, F., Horrocks, I., McGuinness D.L., Patel-Schneider, P.F. 2001. OIL: An ontology infrastructure for the semantic web, *IEEE Intelligent Systems, 16(2):38–45, 2001*

[Franck99]
Franck, G. 1999. Jenseits von Geld und Information. Zur Ökonomie der Aufmerksamkeit, *medien+erziehung, 43. Jahrgang, Heft 3/1999, München, S.146-153.*

[Goldhaber97]
Goldhaber, M. 1997. The Attention Economy and the Net, *(2nd) Draft version of a talk presented at the conference on "Economics of Digital Information," Cambridge, MA, Jan. 23-26, 1997*
(http://www.well.com/user/mgoldh/AtEcandNet.html, zuletzt besucht am 23.02.2005)

[Goorhuis94]
Goorhuis, H. 1994. Konstruktivistische Modellbildung in der Informatik, *Universität Zürich, Dissertation*

[HampFeld97]
Hamp, B., Feldweg, H. 1997. GermaNet: a Lexical-Semantic Net for German, *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications". Madrid, 1997*

[Hanneman01]
Hanneman, R.A. 2001. Introduction to Social Network Methods, *Department of Sociology, University of California*

[HoPatHar02]
Horrocks, I., Patel-Schneider, P. F., van Harmelen, F. 2002. Reviewing the Design of DAML+OIL: An Ontology Language for the Semantic Web, *Proceedings of the 18th Nat. Conference on Artificial Intelligence*

[HoPatHar03]
Horrocks, I., Patel-Schneider, P. F., van Harmelen, F. 2003. From SHIQ and RDF to OWL: The Making of a Web Ontology Language, *Journal of Web Semantics*

[HoSattTob00]
Horrocks, I., Sattler, U., Tobies S. 2000. Practical Reasoning for Very Expressive Description Logics, *Logic Journal of the IGPL 8(3):239-264, May 2000*

[ISO13250]
International Organization for Standardization. 2000. ISO/IEC 13250, Information technology - SGML Applications - Topic Maps, *Genf*

[ISO2788]
International Organization for Standardization. 1996. ISO 2788, Documentation – Guidelines for the establishment of monolingual thesauri, *2. edition. 1986-11-15, Reconfirmed 1996, Genf*

[Jansen99]
Jansen, D. 1999. Einführung in die Netzwerkanalyse: Grundlagen, Methoden, Anwendungen, *Leske + Budrich, Opladen*

[JurMartin00]
Jurafsky, D., Martin, J.H. 2000. Speech and Language Processing, an introduction to natural language processing, computational linguistics, and speech recognition, *Prentice Hall, Upper Saddle River, NJ*

[Kunze01]
Kunze, C. 2001. Lexikalisch-semantische Wortnetze, *In: Carstensen, K.-U. et al. (Hg.): Computerlinguistik und Sprachtechnologie: eine Einführung*, *Spektrum Akademischer Verlag, Heidelberg, Berlin, S. 386-393.*

[LauZieg04]
Lausen, G., Ziegler, C.-N. 2004. Spreading Activation Models for Trust Propagation, *IEEE International Conference on e-Technology, e-Commerce, and e-Service (EEE '04), March 29-31, 2004, Taipei, Taiwan*

[MaSchlie02]
Malsch, T., Schlieder C. 2002. Communication without Agents? From Agent-Oriented to Communication-Oriented Modeling, *Regulated Agent-Based Social Systems: First International Workshop, RASTA 2002, Bologna, Italy, July 16, 2002, Revised Selected and Invited Papers. Springer-Verlag, Berlin, Heidelberg, New York, S. 113-133.*

[Miller&al90]
Miller, G. [u.a.] 1990. Five Papers on WordNet, *Journal of Lexicography, Bd 3(4), S. 235-312, 1990*

[Miller&al95]
Miller, G. [u.a.] 1995. WordNet: A Lexical Database for English, *Communications of the ACM, Vol. 38, Issue 11, Nov. 1995, S. 39-41*

[Miller98]
Miller, E. 1998. An Introduction to the Resource Description Framework, *D-Lib Magazine, May 1998, Dublin, Ohio*

[MüWid02]
Mück, T., Widhalm, R. 2002. Topic Maps: Semantische Suche im Internet, *Springer-Verlag, Berlin [u.a.]*

[NeiEilPfe98]
Neidhardt, F., Eilders, C., Pfetsch, B. 1998. Die Stimme der Medien im politischen Prozeß: Themen und Meinungen in Pressekommentaren, *Discussion Paper FS III 98-106. Wissenschaftszentrum Berlin (WZB)*

[OWLGuide04]
Smith, M.K., Welty, C., McGuinness, D. (eds.) 2004. OWL Web Ontology Language Guide, *W3C Recommendation 10 February 2004*
(http://www.w3.org/TR/2004/REC-owl-guide-20040210/, zuletzt besucht am 23.02.2005)

[OWLRef04]
Dean, M., Schreiber, G. (eds.) 2004. OWL Web Ontology Language Reference, *W3C Recommendation 10 February 2004*
(http://www.w3.org/TR/2004/REC-owl-ref-20040210/, zuletzt besucht am 23.02.2005)

[Page&al98]
Page, L., Brin, S., Motwani, R., Winograd, T. 1998. The PageRank Citation Ranking: Bringing Order to the Web, *Stanford Digital Libraries Working Paper*

[Pepper00]
Pepper, S. 2000. The TAO of Topic Maps. Finding the Way in the Age of Infoglut, *Proceedings of XML Europe 2000, Paris, France*

[Preece81]
Preece, S. 1981. A spreading activation network model for information retrieval, *PhD thesis, CS Dept., Univ. of Illinios, Urbana, IL.*

[Quillian68]
Quillian, R. 1968. Semantic Memory, *in: Minsky, M. (ed), Semantic Information Processing, pages 227–270. MIT Press, Boston, CA, USA, 1968.*

[RDFSchemaSpec1.0]
Brickley, D., Guha, R.V. (eds.) 2000. Resource Description Framework (RDF) Schema Specification 1.0, *W3C Candidate Recommendation 27 March 2000*
(http://www.w3.org/TR/2000/CR-rdf-schema-20000327, zuletzt besucht am 23.02.2005)

[RDFSpec99]
Lassila, O., Swick, R. (eds.) 1999. Resource Description Framework (RDF): Model and Syntax Specification, *W3C Recommendation*
(http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/, zuletzt besucht am 23.02.2005)

[Sowa00]
Sowa, J.F. 2000. Knowledge Representation: Logical, Philosophical, and Computational Foundations, *Brooks Cole Publishing Co., Pacific Grove, CA*

[Vossen97]
Vossen, P. 1997. EuroWordNet: a multilingual database for information retrieval, *Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997, Zurich*

[Vossen99]
Vossen, P. (ed.) 1999. EuroWordNet General Document, *EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document*

[Wersig78]
Wersig, G. 1978. Thesaurus-Leitfaden, Eine Einführung in das Thesaurus-Prinzip in Theorie und Praxis, *Verlag Dokumentation Saur KG, München, New York*

[XML1.0]
Bray, T., Paoli, J., Sperberg-McQueen, C.M. 2004. Extensible Markup Language (XML) 1.0 (Third Edition), *W3C Recommendation, 04 February 2004*
(http://www.w3.org/TR/2004/REC-xml-20040204, zuletzt besucht am 23.02.2005)

[XTM1.0]
Pepper, S., Moore, G. (eds.) 2001, XML Topic Maps (XTM) 1.0, *TopicMaps.Org Specification*
(http://www.topicmaps.org/xtm/1.0/, zuletzt besucht am 23.02.2005)

ERKLÄRUNG

Ich erkläre hiermit gemäß §27 Abs. 2 APO, daß ich die vorstehende Diplomarbeit selbständig verfaßt und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

_____ _____

(Datum) (Peter Kiefer)