Facettierte Wissensordnungen und dynamisches Klassieren als Hilfsmittel der Erforschung des Dark Web

Silke Heesemann und Hans-Dieter Nellißen, Düsseldorf

Die Identifizierung und Überwachung terroristischer und/oder extremistischer Aktivitäten und Planungen im Internet, speziell im Dark Web, rücken immer mehr in den Fokus rechtsstaatlicher Ermittlungen. Die Menge an relevanten Informationen erfordert eine Fokussierung auf einen bestimmten Jargon, mit dem verdächtige Entwicklungen erkannt werden. Auf dieser Basis wird ein Analysesystem aufgebaut. Der Versuch eines optimalen Retrieval-Ansatzes zu dieser Thematik wird im Folgenden dargestellt.

Faceted Knowledge Organization Systems and dynamic classification as tools for the research of the Dark Web

The identification and observation of terrorist and/or extremist activities and plans on the internet, especially on the Dark Web, are moving more and more into the centre of inquiry. The amount of relevant information requires to focus on a specific jargon, with the help of which suspicious developments can be detected and an analysis system can be built up. An attempt at an optimal information retrieval approach to this issue is depicted below.

Einleitung

Anfang Mai 2007 wurde das europäische Informationsportal „Check the Web“ bei Europol freigegeben. Im Fokus steht die „Beobachtung und Auswertung des Internets zur Bekämpfung des internationalen Terrorismus“ (BMI, 2007). Das Herausfiltern von zur Terrorismusbekämpfung relevanten Informationen aus dem Internet stellt eine Herausforderung für die Ermittlungsbehörden dar.

Many Internet pages in various languages have to be monitored and evaluated, which requires enormous technical and human resources. Due to the huge quantity of Internet pages in use, problems arise on a national and international level concerning the quantity and quality of resources, especially with a view to the language skills needed (EU Council Secretariat 2007).

Im Dezember 2007 belief sich die Anzahl von Websites im WWW auf über 155 Millionen, mit einem Zuwachs von 50 Millionen seit Dezember 2006, dies entspricht einem prozentualen Anstieg von 48 Prozent (Netcraft Web Server Survey December 2007). Das Interesse der Sicherheitsbehörden und Nachrichtendienste konzentriert sich dabei zusätzlich auf die Überwachung nicht direkt zugänglicher Informationen im Web, so wie Emails, Chatrooms und weiterer Kommunikationswege, die von einem klassischen Crawler nicht abdeckbar sind. Wir befinden uns an dieser Stelle teilweise im Deep Web (Bergman 2001), auch Hidden Web oder Invisible Web genannt (Sherman & Price 2001), das heißt bei Web-Content, der nicht durch Hyperlinks erreichbar ist, wie z.B. Datensätze in Datenbanken. Das Deep Web soll laut einer Studie aus dem Jahr 2000 400- bis 550-mal größer sein als das Oberflächenweb (Bergman 2001). Auch wenn die Zahlen dieser Studie stark angezweifelt werden (Lewandowski 2005), kann man doch von einer großen Menge von Seiten im Deep Web ausgehen. Der Begriff *Dark Web* wird im allgemeinen Sprachgebrauch entweder missverständlich als Synonym für Deep Web verwendet oder er bezeichnet Webseiten mit terroristischem, extremistischem oder anderem kriminellen Inhalt, ob nun im Oberflächen- oder im Deep Web. Unser Thema ist das Web der Terroristen. Zur Überwachung und Informationsgewinnung aus dem Deep Web sind spezielle Techniken nötig, um diese Seiten für einen Crawler überhaupt erreichbar zu machen, da eine einfache Linkverfolgung hier nicht greift. Aufgrund der Komplexität dieses Themas kann an dieser Stelle nicht weiter darauf eingegangen werden.

Die aus Oberflächenweb und Deep Web zusammengesetzte enorme Datenmenge stellt zunächst eine Herausforderung für das Crawlen nach zur Terrorismusbekämpfung relevanten Inhalten dar. Die aus dem Crawlvorgang gewonnene Datenmenge bildet dann die Basis für das Herausfiltern der dort befindlichen Informationen. Es ist daher von Vorteil, ein themenspezifisches, fokussiertes Crawling durchzuführen (Kwiatkowski & Höhfeld 2007) und den Inhalt ebenso thematisch fokussiert durchsuchbar zu machen, um Informationsballast zu vermeiden und qualitativ hochwertige Informationen zu gewinnen. Hierzu benötigt man eine Wissensordnung, die das Focused Crawling und das Retrieval steuert. Doch unabhängig davon wo man nach brisantem Inhalt oder Informationen sucht, wichtig ist die Aufarbeitung des gewonnen Materials in einem Retrievalsystem, welches das Suchen und Finden relevanter Informationen ermöglicht.

Die Untersuchung des Webs konzentriert sich bisher hauptsächlich auf die Aufdeckung und Visualisierung von sozialen Netzwerken, das heißt von Verbindungen zwischen terroristischen Organisationen und deren Mitgliedern sowie deren Identifikation.

Ein Forschungsteam um Hsinchun Chen, Professor of Management Information Systems an der Universität von Arizona, gilt in der wissenschaftlichen Forschung zu diesen Themen momentan als führend (AI Lab 2007). Unter der Leitung von Chen wurde z.B. die Software Coplink entwickelt, welche soziale Netze visualisiert und Ermittlungsbehörden bei der Fahndung nach Straftätern hilft (Chen et al. 2003). Ein weiteres Programm namens Writeprint ermöglicht die Identifizierung anonymer Autoren mit Hilfe von Mustererkennung des Schreibstils und inhaltlicher Analysen, wie die Fokussierung auf Themen und häufig verwendete Schlüsselbegriffe (Li, Zheng & Chen 2006). Writeprint ist Teil des „Dark Web Project“ (Chen et al. 2005). Dieses hat zur Aufgabe, Webseiten extremistischen Ursprungs und Inhalts zu sammeln und auszuwerten (Quin et al. 2005). Hierfür wurden automatisierte Systeme entwickelt:

59(2008)1, 108-117 Information Wissenschaft & Praxis

> To address this research gap, we explore an integrated approach for identifying and collecting terrorist/ extremist Web contents. We also propose a Dark Web Attribute System (DWAS) to enable quantitative Dark Web content analysis from three perspectives: technical sophistication, content richness, and Web interactivity (Qin et al. 2007).

Neben der Aufdeckung sozialer Netze wird also auch die technische und inhaltliche Entwicklung extremistischer Webseiten erforscht. Das Attribute System basiert auf den Untersuchungen Demchaks der ähnliche Untersuchungen für Regierungs-Webseiten unternommen hat (Demchak 2000; Zhou 2005). Demchaks System wurde unter anderem um die Attribute Propaganda und Ideologie erweitert; es werden aber nur Cluster von Webseiten identifiziert, in denen entsprechende Begriffe auftreten (Reid et al. 2005). Der Fokus dieser Untersuchungen liegt folglich nicht in der inhaltlichen Erschließung der gewonnenen Dokumente. Dieser Punkt der inhaltlichen Analyse wurde bisher selten beachtet. Die steigende Anzahl von Dokumenten, die gewonnen wird, muss von den Ermittlungsbehörden aber auch inhaltlich effizient durchsuchbar gemacht werden. Die Datensammlung des „Dark Web Project“ beinhaltet 500 Millionen Dokumente (Webseiten, Postings, Videos etc..) von 10.000 Websites; eine Datengröße von insgesamt zwei Terabyte (AI Lab, 2007). Das zugehörige „Dark Web Research Portal“ bietet zwar eine Suchoberfläche an, mit welcher nach Schlüsselwörtern im Datenbestand gesucht werden kann, allerdings liegt hier keine komplexe Ontologie zugrunde (Reid & Chen 2006). Die Möglichkeiten, die das Information Retrieval durch den Einsatz semantischer Netze in Verbindung mit facettierten Begriffsordnungen bietet, werden nicht genutzt. Gerade hier liegt ein Schwerpunkt des Instituts für Sprache und Information, Abteilung Informationswissenschaft der Heinrich-Heine-Universität Düsseldorf.

Im Rahmen eines Projekts der Abteilung Informationswissenschaft an der Heinrich-Heine-Universität Düsseldorf wurde ein exemplarischer Thesaurus über den Terror in Nordirland erstellt und mit Hilfe des Retrievalsystems RetrievalWare der Firma Convera realisiert (RetrievalWare wurde mittlerweile vom Wettbewerber FAST übernommen).

Mit Hilfe facettierter Thesauri und dynamischer Klassierung, wie sie die Software RetrievalWare ermöglicht, können Wissensdomänen unter Einsatz eines semantischen Netzes abgebildet und Datenbasen unterschiedlicher Formate und Sprachen nach Informationen durchsucht werden. Mit dem in RetrievalWare integrierten Tool Knowledge Workbench kann man ein kontrolliertes Vokabular in Form einer Taxonomie erstellen, die dann durch Begriffserweiterungen in Form von Relationen ein semantisches Netz bildet. Die Software ermöglicht die Erstellung eines facettierten, multilingualen Thesaurus, auf dessen Grundlage eine dynamische Klassierung der Dokumente aus der Datenbasis erfolgt und dann dem Nutzer in der Suchoberfläche vielfältige Recherchemöglichkeiten bietet. Dreh- und Angelpunkt dieses Retrievalsystems ist der Thesaurus oder besser gesagt, sind die Thesauri.

Facettierte Wissensordnungen

Der Einsatz von Facetten zur Inhaltserschließung ist nicht neu (Stock & Stock 2008). Bereits 1930 stellt Ranganathan sein System facettierter Klassifikation vor (Ranganathan 1933). Die von ihm entwickelte Colon-Classification ordnet Themen in die Facetten Personality (Individualität), Matter (Material), Energy (Energie), Space (Raum) und Time (Zeit) ein. In einer Facettenklassifikation wird ein Gegenstand oder Thema hiermit aus unterschiedlichen Fragestellungen zu seiner Beschaffenheit oder des thematischen Inhalts klassifiziert und fällt somit mehreren Klassen zu. Das Thema oder der Gegenstand wird aus unterschiedlichen Dimensionen betrachtet:

> In short, faceted classification is a method of multidimensional description and arrangement of information resources by their concept, attributes or "aboutness". It addresses the fact that users may look for a document resource from any number of angles corresponding to its rich attributes (Uddin & Janecek 2007).

Sucht man z.B. in einer Datenbank nach einem Artikel über Architektur im 19. Jahrhundert, lässt sich die Suche in einer Vorauswahl auf den entsprechenden Zeitrahmen oder eine Stilrichtung einschränken, und es werden nur die Dokumente angezeigt, die diesen Einschränkungen entsprechen. Allerdings werden die von Ranganathan aufgestellten starren Regeln zur Facetten-Erstellung im Allgemeinen nicht streng übernommen. Die Facetten werden frei zum entsprechenden Thema passend erstellt, wobei genau diese Flexibilität ein großer Vorteil der Facetten-Klassifikation darstellt. Jede Facette kann einer eigenen Struktur in der Klassifikation folgen. So kann z.B. eine materielle Facette als Hyponomie hierarchisch aufgebaut sein, während eine räumliche Facette als Meronymie, das heißt Teil-Ganzes Beziehung aufgebaut ist (Kwasnik 1999). Die Kategorien einer facettierten Klassifikation unterliegen keinem starren enumerativen Schema, sondern sind frei untereinander kombinierbar (Vickery 1969). Der erste Schritt bei der Erstellung einer facettierten Wissensordnung ist die Facettenanalyse, an deren Ende eine Liste von Kategorien steht, welche die Facetten bilden (Broughton 2006).

Heute sind facettierte Klassifikationen im Web häufig zu finden, da sie die Navigation durch den Content erleichtern. So bieten viele Webshops die Möglichkeit, unter verschiedenen Spezifikationen nach Waren des Sortiments zu suchen. In der Suchoberfläche lässt sich die Artikelauswahl z.B. nach Produktart, Preisrahmen und/oder Hersteller durch ein Dialogfeld oder ein Verzeichnis einschränken. Die Eingabe eines Suchworts ist dabei kaum noch nötig. An der University of California in Berkeley wurde die Open-Source-Software „Flamenco“ (FLexible information Access using MEtadata in Novel COmbinations) entwickelt, mit der man eine Suchoberfläche erstellen kann, die Dokumente in Kategorien einordnet, durch die Nutzer dann durch Hyperlinks in den Verzeichnissen browsen können (Elliott 2001). Diese flexible browsing-basierte Suche in kategorisierten Verzeichnissen erwies sich in Nutzerstudien als vorteilhaft (Hearst et al. 2002; Yee et al. 2003).

Dem Nutzer muss keine Suchterminologie bekannt sein, um komplexe Anfragen zu stellen. Er kann seine Suche schrittweise nach seinen Ansprüchen einschränken oder erweitern (English et al. 2002a). Wählt man eine Kategorie aus einer Facette aus, so werden die Ergebnisse aus dieser angezeigt, sowie eine Navigationsleiste mit den anderen Kategorien, die ebenfalls mit den Ergebnissen übereinstimmen und weitere Unterkategorien der ersten Auswahl. Die Suche kann so immer wieder spezifiziert und um andere Aspekte erweitert oder wieder abstrahiert werden, ohne dass der Nutzer den Überblick verliert (English et al. 2002b). Eine oft frustrierende Versuch-und-Irrtum-Interaktion mit dem Suchsystem ist somit nahezu ausgeschlossen (Sacco 2006).

Für Produkte und Bilder müssen in den jeweiligen Dokumenten Metadaten (Tags) hinzugefügt werden, die das Objekt entsprechend den Kategorien beschreiben. Bei der Suche innerhalb von Texten bestehen diese Metadaten aus dem Text selbst. Dokumente, in denen ein Wort vorkommt, welches mit einer Kategorie übereinstimmt, werden als Ergebnis angezeigt. In komplexen Retrievalsystemen geschieht dies jedoch nicht allein durch eine thematische Volltextindexierung, sondern durch ein semantisches Netz.

Facettierter Thesaurus

Ein Dokumentationsgebiet bzw. Wissensgebiet wird durch eine Begriffsordnung dargestellt, die dieses Wissensgebiet abbildet. Es wurden bereits vielfältige Thesauri unterschiedlichster Wissensgebiete erarbeitet, wie z.B. der umfassende Standard-Thesaurus Wirtschaft.
Bei einem facettierten Thesaurus wird ein Wissensgebiet ebenso begrifflich dargestellt, allerdings aus mehreren Blickwinkeln betrachtet. Dem Nutzer steht in einer entsprechenden Suchoberfläche nicht nur eine einzige Begriffsordnung zur Verfügung, sondern gleich mehrere, die untereinander kombinierbar sind. Die facettierte Klassifikation wird durch Relationen erweitert, so dass ein facettierter Thesaurus entsteht. Ein früher facettierter Thesaurus ist der „Thesaurofacet" des Wissensgebiets Technik und verwandten Themen (Aitchison, Gomershall & Ireland 1969). Der Nachrichtendienst Factiva arbeitet mit den vier Facetten Unternehmen, Branchen, Geographica und Themen, allerdings ausschließlich mit Hierarchierelationen (Stock 2002). Es ist aber möglich, noch komplexere multilinguale facettierte Ontologien umzusetzen.

RetrievalWare arbeitet bei der Suche nach Informationen mit semantischen Netzen. Ein Thesaurus stellt so ein semantisches Netz dar. Begriffe eines Themengebiets werden systematisiert und mit einer Vorzugsbenennung, dem Deskriptor, in einer Baumstruktur dargestellt. Synonyme verweisen auf den entsprechenden Deskriptor. Weitere Relationen, wie z.B. Ober- und Unterbegriffe, bilden das Konzept um den jeweiligen Begriff ab. Diese Verbindung von Konzepten mittels Relationen bildet das semantische Netz. Es kann jedoch nicht nur ein einziger Thesaurus erstellt und in RetrievalWare integriert werden, sondern es sind gleich mehrere Thesauri zum Indexieren, Klassifizieren und Suchen miteinander kombinierbar. Es entsteht so ein facettierter Thesaurus, der dann (unter ausschließlicher Nutzung der Hierarchierelation) als facettierte Klassifikation eine facettierte Suche nach Informationen ermöglicht.
Für den Beispielthesaurus „Terror in Nordirland" musste zunächst überlegt werden, welche Facetten das zu erfassende Themengebiet umfasst. Hierzu war eine Recherche nötig, um sich mit dem Thema vertraut zu machen. Im Rahmen des zeitlich begrenzten Projekts konnte natürlich nur ein rudimentärer Einblick in das Themengebiet gewonnen werden. Es musste überlegt werden, welche Aspekte für die Begriffsordnung relevant sind. Es wurden drei Thesauri erstellt: Ein Thesaurus beinhaltet geographische Angaben, die zum Teil bis auf die Straßenebene führen oder spezielle Lokalitäten beinhalten, wie z.B. die Haftanstalt Maze Prison, in der viele Mitglieder von terroristischen Organisationen inhaftiert waren. Ein weiterer Thesaurus stellt Organisationen zusammen, sowohl Terrororganisationen als auch Regierungsorganisationen, denen zum Teil Personen zugeordnet wurden. Der dritte Thesaurus befasst sich mit allgemeinen, themenrelevanten Begriffen, wie z.B. Waffen, terroristischen Aktivitäten und bereits vorgefallenen Anschlägen, im Projekt „misc" genannt.
Es werden somit drei Fragestellungen abgedeckt: Wer? (Organisationen), Wo? (Geographie) und Was? (misc). Nach Ranganathan wären dies die Facetten Personality, Space und Matter. Dem Nutzer können in diesem Beispiel drei Thesauri zur Verfügung gestellt werden, die individuell kombinierbar sind und so eine Sicht auf verschiedene Aspekte des Terrorismus aus unterschiedlichen Perspektiven eröffnen.

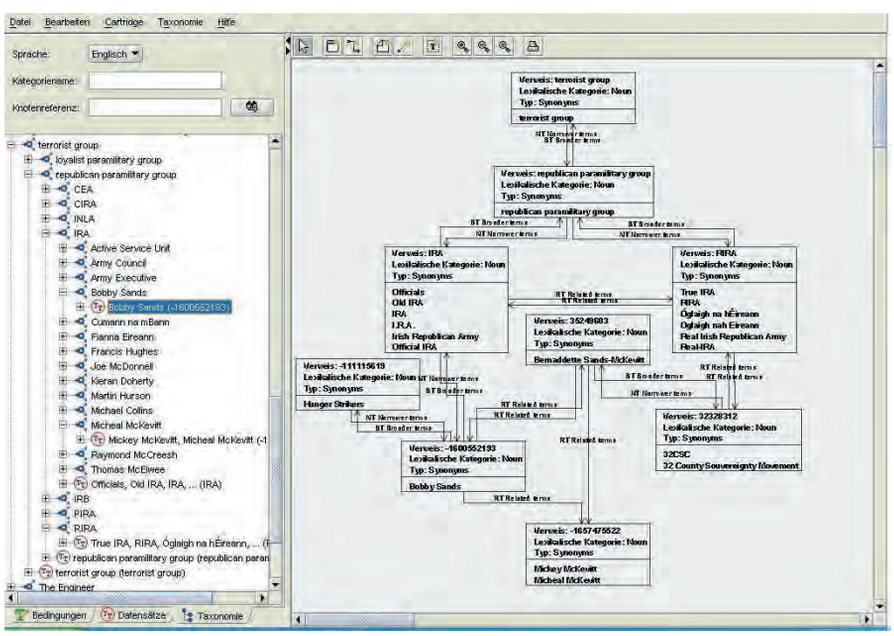

Abbildung 1: CartridgeEditor Convera.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

Ein Dokument, welches einen Anschlag der IRA in Belfast thematisiert, wird somit jeder Thesaurusfacette zugeordnet und ist aus jeder Facette erreichbar.
Der Einsatz facettierter Thesauri ermöglicht zudem eine schnelle Einsatzfähigkeit. So können die Ermittler schon mit wenigen kleinen Facetten arbeiten und das Retrievalsystem nach und nach um weitere Facetten erweitern und diese mit steigendem Hintergrundwissen und neuen Informationen modifizieren. Allerdings ist dabei zu beachten, dass ein bereits vorhandener Datenbestand bei einer Änderung bzw. Erweiterung der facettierten Thesauri erneut indexiert und klassiert werden muss.
Der oft kritisierte Mangel an Flexibilität und Skalierbarkeit von Ontologien wird durch den Einsatz facettierter Thesauri aufgehoben. Die Facetten gewähren einen leichten Überblick und können flexibel modifiziert werden. Ein einzelner Thesaurus mit 10.000 Begriffen kann so gegebenenfalls auf zwei Facetten mit je 100 Begriffen reduziert werden. Die Möglichkeit der Kombination dieser beiden Facetten, also 100·100, ergibt dann ebenfalls 10.000 Begriffe (Buchanan 1989).

Thesauruserstellung mit Convera

Begonnen wird mit der Erstellung der Taxonomie, das heißt einer monohierarchischen Klassifikation von Begriffen, welche in einer Baumstruktur dargestellt wird. Convera bietet zur Erstellung der Begriffsordnung die Software Knowledge Workbench an. Mit Hilfe des darin integrierten Programms Cartridge Editor kann ein komplexes Begriffsnetz erstellt werden. Dieses Begriffsnetz dient dann zur Indexierung, Klassierung und Gewichtung der Dokumente und zur Filterung der gesuchten Informationen innerhalb der Datenbasis. Es können aber auch bereits zur Verfügung stehende Thesauri in RetrievalWare eingebunden werden (Bayer et al. 2005).
Der Editor hat zwei Bereiche, wie in Abbildung 1 zu sehen ist. Im linken Bereich gibt es drei Registerkarten – Bedingungen, Datensätze und Taxonomie –, welche die terminologische Kontrolle unterstützen. Im rechten Bereich befindet sich eine graphische Oberfläche, Canvas genannt, in welcher die Beziehungen zwischen den Begriffen erstellt und visualisiert werden. Auf dem Register „Taxonomie" werden die Kategorien monohierarchisch in einer Baumstruktur erstellt. Zu jeder Kategorie wird automatisch ein Datensatz generiert, von Convera Synonym Set, kurz Synset genannt.

In Abbildung 2 ist ein solches Synset für den Begriff *IRA* aus der Thesaurusfacette „Organisationen“ zu sehen.
Beim Erstellen der Kategorie ist der gewählte Kategoriename in der Taxonomie

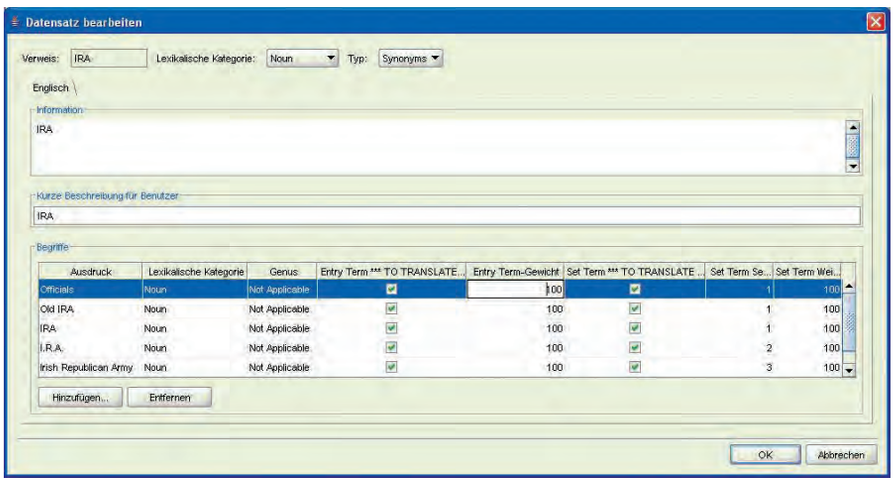

Abbildung 2: Synset für den Begriff IRA aus der Thesaurusfacette „Organisationen“. Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

die Vorzugsbenennung, also der Deskriptor, welcher später in der Suchoberfläche dem Nutzer als Verzeichnis angezeigt wird. Diesem Deskriptor können Synonyme zugefügt und deren grammatikalische Kategorie und Genus angegeben werden. Des Weiteren kann für jedes Synonym ein Gewichtungswert gesetzt werden, der den Bindungsgrad der Bezeichnung an den Begriff angibt und beim Ranking der Dokumente dementsprechend berücksichtigt wird. Würde der Begriff *Old IRA* mit einem Term-Gewicht von 50 Prozent angegeben, würde dieser bei der Klassifikation und bei der Suche durch den Nutzer auch nur mit 50 Prozent in das Ranking eingehen. Bei dem gewählten Synonym-Gewicht von 100 Prozent ergibt die Suchanfrage nach *IRA* z.B. die gleichen Ergebnisse wie die Suche nach dem hinterlegten Synonym *Irish Republican Army*.
Es muss nun jede Kategorie intellektuell und manuell erstellt werden, wobei eventuell weitere Datensätze heranzuziehen sind, um das semantische Umfeld des Begriffs genau abzubilden und die Relevanz der Treffer bei der Suche zu erhöhen. Bei der Erstellung der Taxonomie muss bedacht werden, dass diese dem Nutzer später als Verzeichnis in der Suchoberfläche angezeigt wird.

> Es (...) ist dafür Sorge zu tragen, dass pro Begriff eine überschaubare Menge an Unterbegriffen (etwa maximal 20 bis 30) existiert, da ansonsten die Ausgabe unübersichtlich werden kann. Dasselbe gilt für die Anzahl der Topterme der Dokumentationssprache (Stock 2007, 457).

Bei der Erstellung der Synsets muss ein sehr gutes Verständnis des Themas vorhanden sein. So ist z.B. Robert Bates, ein Mitglied der „Shankill Butchers“, einer besonders brutalen Gruppe der „Ulster Volunteer Force“ Anfang der 70er Jahre, auch unter dem Namen „Basher“ bekannt. Es ist wichtig solche Pseudonyme zu kennen, weil sonst wichtige Dokumente nicht gefunden werden können. So erhält man für die Suchanfrage *Basher* alle Dokumente, die den Namen Robert Bates enthalten, und umgekehrt alle Dokumente, die Bates nur unter seinem Pseudonym behandeln. Durch die Eingabe des Synonyms *Basher* in den Datensatz wird dem Nutzer der Zugang zu diesen Dokumenten ermöglicht, ohne dass er alle Pseudonyme kennen muss. Im Registerblatt „Datensätze“ können alle Synsets sortiert in der Reihenfolge ihrer Eingabe angezeigt werden. Das Programm weist jedem Datensatz automatisch eine Verweisnummer als Identifikator zu; sie kann aber auch manuell eingegeben werden. Im Register „Bedingungen“ werden alle Worte, sowohl die Vorzugsbenennungen als auch die Nicht-Deskriptoren der Synsets, alphabetisch aufgelistet. Nachdem alle Begriffe eingegeben wurden, kann zu der Herstellung der Relationen im Canvas übergegangen werden. Es können beliebige Relationsformen mit unterschiedlichen Gewichtungswerten hergestellt werden. Für die Thesauri des Projekts wurden die Relationen Broader Term (BT), Narrower Term (NT) und Related Term (RT) eingesetzt (Abbildung 1).
Die manuelle Eingabe erweist sich als aufwendig und erfordert ein hohes Maß an Präzision und Kontrolle. Die erstellten Taxonomien und Thesauri können mit Hilfe von Qualitätsmetriken und eines Benchmarking auf ihre Effizienz getestet und modifiziert werden. Nach positiver Evaluation wurden die drei Thesaurusfacetten in RetrievalWare eingebunden (Abbildung 3).

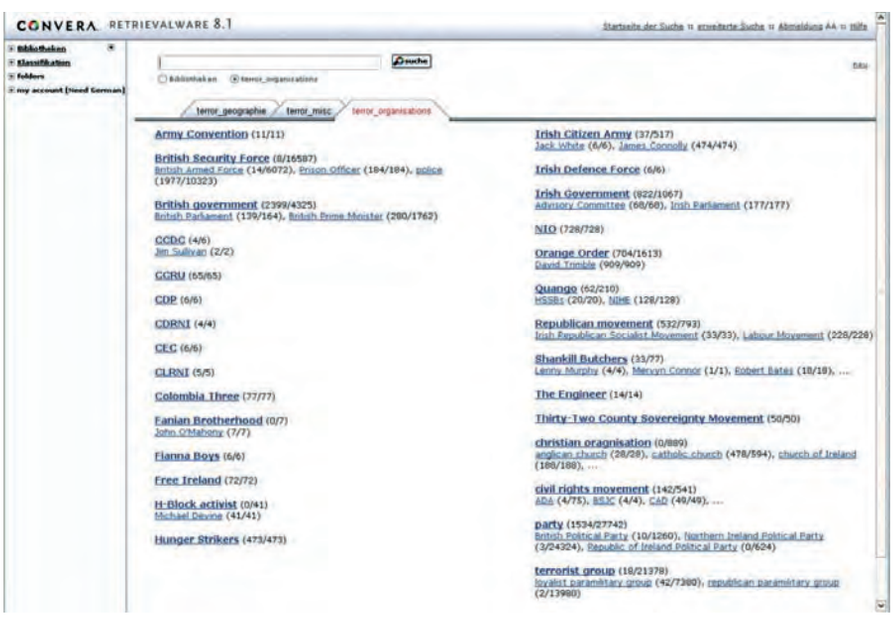

Abbildung 3: Ansicht Startseite der Suche. Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

Crawling

Nachdem die drei Thesauri fertig gestellt waren, musste eine Datenbasis aus dem Web gewonnen werden. Da der Crawler von Convera nicht lizenziert war, griff die Projektgruppe auf ein Freeware-Programm unter GPL-Lizenz zurück. Ein Crawler, auch Robot oder Spider genannt, verfolgt Links, die sich auf einer Webseite befinden und speichert deren Inhalte ab (Pant, Srinivasan & Menczer 2003). Hierzu muss zunächst eine Liste mit relevanten Start-URLs, auch Seed-Liste genannt, vorhanden sein, von der aus der Crawler die dort hinterlegten Hyperlinks verfolgt. Die Seed-Liste mit den Ausgang-URL

unseres Projekts umfasste unter anderem frei zugängliche Webseiten von Organisationen (Sinn Fein), Tageszeitungen (An Phlobacht) und Datenbanken (Cain), von denen auch relevante weiterführende Links zu erwarten waren. Es ist dabei zu beachten, dass die Tiefe der Linkverfolgung nicht zu hoch eingestellt wird, weil sonst zu viele irrelevante Seiten gespiegelt werden. Der Crawler wurde so konfiguriert, dass er alle Datei-Formate übernommen hat. Insgesamt wurden für die Experimentelle IRA-Datenbank 1,5 Gigabyte an Daten zusammengestellt.

Unter realen Bedingungen, das heißt bei einem Vorhaben wie „Check the Web" wäre es in Anbetracht der vorhersehbaren enormen Datenmenge von Vorteil, einen Crawler zu verwenden, der fokussiert auf ein Thema hin crawlt (Kwiatkowski & Höhfeld 2007).

Dynamische Klassierung und facettierte Thesauri

Die Dynamische Klassierung ermöglicht das Indexieren und Kategorisieren in einem Arbeitsschritt. So kann auch eine sehr große heterogene Datenbasis, wie sie sich aus einem Crawling ergibt, schnell und präzise durchsucht werden.
Die dynamische Klassierung erfolgt auf Grundlage des erstellten facettierten Thesaurus. Die Dokumente werden vom Retrievalsystem mit den Termen der Thesauri indexiert und nach ihrer Gewichtung bewertet. In der Suchoberfläche werden die erstellten Taxonomien mit der Trefferzahl zum jeweiligen Begriff und dessen untergeordneten Begriffen als Verzeichnisse angezeigt. Die indexierten Dokumente werden dann direkt den Klassen zugeordnet (Abbildung 4).

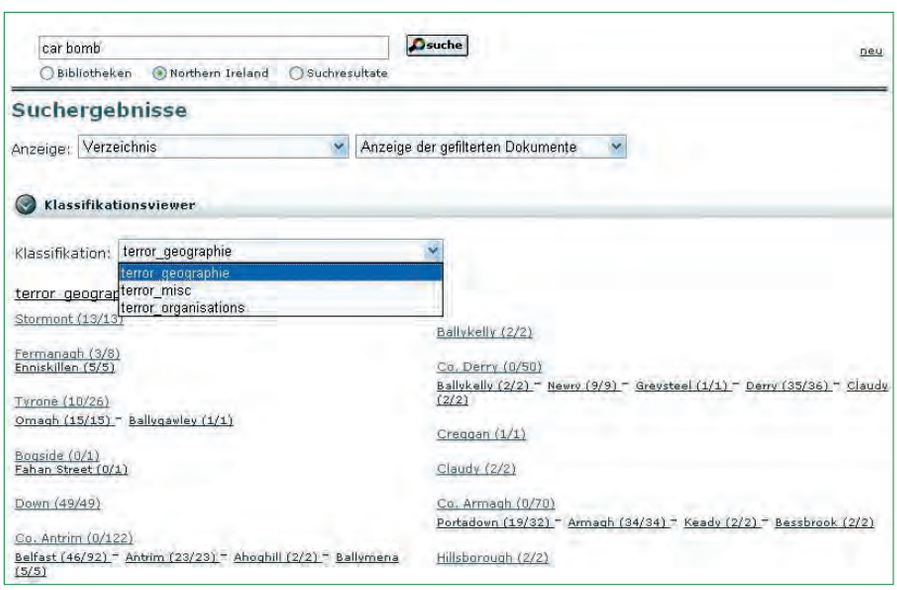

Abbildung 4: Verzeichnis-Ansicht Suchergebnisse.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft

Das Focused Crawling ermöglicht das Durchsuchen des *Dark Webs* auf der Grundlage einer Ontologie (Chakrabarti, van den Berg & Dom 1999). Ein fokussierter Crawler findet durch eine Begriffsordnung effektiv und ballastarm thematisch verwandte Seiten (Davison 2000). Das Problem einer zu hohen Link-Tiefe wird durch das fokussierte Crawling gelöst, da nur relevante Seiten verfolgt werden. Eine Beschreibung dieser sinnvollen Vorgehensweise würde den Rahmen dieses Artikels sprengen, sollte in der Praxis aber weiter verfolgt und umgesetzt werden.
Um die Aktualität des gecrawlten Datenbestandes zu gewährleisten, muss der Crawlvorgang in kurzen Abständen immer wieder neu gestartet werden, denn ein großer Teil der aktiven Web-Inhalte verändert sich täglich (Cho & Garcia-Molina 2000).

Die Suche nach *car bomb* in der Thesaurusfacette „Geographie" führt für den Begriff *Co. Antrim* zu keinem Treffer, allerdings zu 122 Treffern in den Unterbegriffen, also zu den Stadtteilen, die zu County Antrim gehören. Darunter sind wiederum 46 Treffer für den Begriff *Belfast* und 92 Treffer für dessen Unterbegriffe, wie z.B. *Maze Prison*. Es ist also möglich im Verzeichnis nach den Suchergebnissen zu browsen und so immer tiefere, das heißt speziellere Fokusse zu betrachten. Die Begriffsleiter, die der Nutzer beim Browsen durch die Begriffe hinter sich gelassen hat, wird dabei in jedem Schritt angezeigt, damit er sich besser orientieren und den Pfad seiner Suche zurückverfolgen kann. In unserem Beispiel in Abbildung 4 hat der Nutzer die Suche nach dem Begriff *car bomb* in der Vorauswahl auf die Facette „Geographie" mit Fokus auf *Northern Ireland* eingeschränkt. Dieselbe Suchanfrage lässt sich allerdings auch innerhalb dieses Fokus auf die anderen Facetten anwenden, indem im Klassifikationsviewer eine andere Facette ausgewählt wird. Die angezeigten Treffer zeigen dann z.B. bei gewählter Facette „Organisationen", die Dokumente an, die den Begriff *car bomb* sowohl im Zusammenhang mit dem Begriff *Northern Ireland*, als auch mit den Begriffen der Organisationen nennen. So ergänzen sich dynamische Klassierung und facettierte Thesauri zu einfach durchführbaren und zugleich komplexen Anfragen.
Die Anzeige der Suchergebnisse in der Tabellenansicht, wie sie Convera RetrievalWare anbietet, stellt dabei eine besonders nutzerfreundliche Visualisierung der Suchergebnisse dar und ermöglicht die dynamische Anzeige von Treffern in allen drei erstellten Facetten.

In Abbildung 5 hat der Nutzer am Beginn seiner Suche den Begriff *Belfast* im Verzeichnis „Geographie" gewählt und den Suchbegriff *body count* eingegeben. In der Tabellenansicht hat er dann die Möglichkeit, zwei weitere Facetten einzubeziehen. In diesem Beispiel steht horizontal die Begriffsleiter der Organisationen und vertikal die Begriffsleiter der allgemeinen Begriffe (misc) zur Verfügung. Auch hier wird der Pfad, den der Nutzer innerhalb der Begriffsleiter genommen hat, angezeigt. In der Facette „Organisationen" befindet er sich auf der Ebene der republikanisch paramilitärischen Gruppierungen und in der Facette der allgemeinen Begriffe auf der Ebene der verschiedenen Arten von Bomben. In der Tabelle wird die Anzahl Treffer zu dem Begriff *body count* im Kontext dieser Begriffsebenen angezeigt.
Eine einzige Suchanfrage kann hiermit auf Konzepte erweitert werden, die die Treffermenge signifikant auf den zu betrachtenden Fokus eingrenzen und Informationsballast vermeiden. Die 25 Dokumente, die im dargestellten Beispiel unter *IRA* und *car bomb* angezeigt werden, enthalten Begriffe, die im Kontext zu Todesopfern im Zusammenhang mit der Organisation IRA und Autobomben im geographischen Gebiet Belfast stehen. Dieses Beispiel zeigt eine extrem fokussierte Ansicht, die jedoch jederzeit wieder verallgemeinerbar ist, indem der Nutzer sich in der Begriffsleiter wieder nach oben bewegt. Die Tabellen-Ansicht der dynamischen Klassierung hat jedoch nicht nur den Vorteil, dass die Treffermenge kontextspezifisch noch mehr eingegrenzt werden kann als in der einfachen Verzeichnisstruktur, sie bietet auch die Möglichkeit, Muster und Verbindungen aufzudecken, die bei der initialen Suchanfrage noch gar nicht in Betracht gezogen worden sind. Die Granularität des semantischen Netzes lässt sich so an-

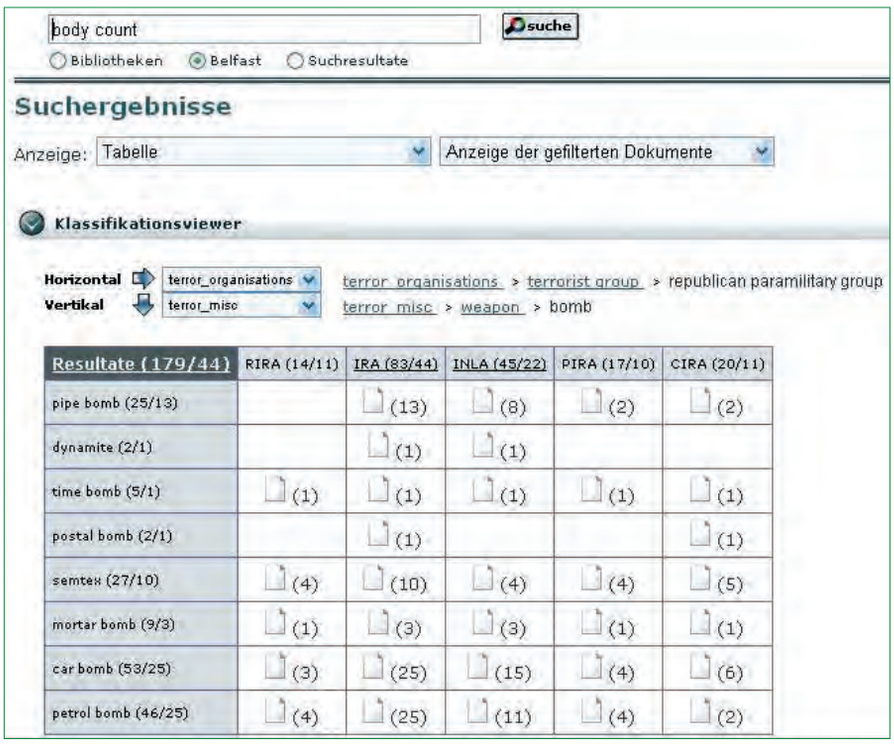

Abbildung 5: Tabellen-Ansicht Suchergebnisse.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

fragespezifisch einstellen, dass entweder fokussiert oder abstrahiert wird, wobei eine Abstraktion Informationen aufdecken kann, die aus einer Fokussierung herausgefallen sind und so unerwartet Fragen beantworten kann, die noch gar nicht gestellt wurden.

Ranking und Ergebnisansicht

Die Gewichtung (und in deren Folge das Ranking) der Dokumente erfolgt bei Convera auf unterschiedlichen Ebenen. Die eingegebenen Suchterme werden mit den indexierten Dokumenten abgeglichen und unter verschiedenen Aspekten gewichtet. Der Rankingprozess beginnt mit einem grobkörnigen Ranking (coarse grain rank) und wird mit einem feinkörnigen Ranking (fine grain rank) abgeschlossen. Die aus jedem Schritt resultierenden Gewichtungswerte für jedes Dokument werden addiert. Das erste grobkörnige Ranking gewichtet die Dokumente nach der Anzahl der mit den in der Suchanfrage übereinstimmenden Terme und deren verwandten Begriffen. Danach werden die verwandten Begriffe danach gewichtet, in welchem Kontext sie stehen, das heißt nach ihrer Nähe zu anderen kontextspezifischen Begriffen. Die Begriffe werden auch nach ihrem Verwandtschaftsgrad gewichtet, das heißt stark verbundene Konzepte werden höher gewichtet als schwach verbundene Konzepte. Morphologische Variationen werden dabei ebenfalls berücksichtigt. Verfeinert wird das Ranking durch das „fine grain". Je näher die Terme der Suchanfrage im Dokument zusammen auftreten, desto höher ist der Gewichtungswert. Kommen zwei Suchterme oder deren verwandte Begriffe in einem Satz vor, so ist das Dokument relevanter als wenn diese im selben Absatz vorkommen. Diese Stellungsrestriktion macht auch eine Disambiguierung von Homonymen möglich. Des Weiteren werden die Anzahl der gefundenen Suchterme und deren Entsprechungen in Relation zur Länge des Dokuments gestellt. Das grobkörnige Ranking lässt sich bei der Erstellung der Thesauri durch einstellbare Gewichtungswerte der verschiedenen Relationen beeinflussen. Das feinkörnige Ranking kann vom Administrator vor dem Indexlauf eingestellt werden.
Die Ergebnisse aus der Suchanfrage werden dem Nutzer mit der Angabe der Dokumentrelevanz in einer Liste angezeigt. Neben einer Ansicht der Treffer mit einer kurzen Schlagzeile der Dokumente steht auch die Ansicht einer Dokumentzusammenfassung zur Verfügung, in welcher der thematische Schwerpunkt des jeweiligen Dokuments ersichtlich ist (Abbildungen 6 und 7).

Wählt man ein Dokument aus der Ergebnisliste aus, so wird dieses mit den markierten Treffern angezeigt (Abbildung 8). Dabei werden die gefundenen Begriffe aus der Suchanfrage gelb und die verwandten Begriffe blau markiert. Bewegt man den Cursor auf einen markierten Begriff, so wird der errechnete Relevanzwert für diesen angezeigt.

Darüber hinaus können die Suchterme vom Nutzer manuell gewichtet werden, indem die jeweiligen Wörter mit einem Doppelpunkt (Multiplikator) und einer Zahl von eins bis zehn versehen werden. So würde bei einer Suchanfrage mit dem Ausdruck *bomb Engineer:10* der Term *Engineer* zehnmal höher gewichtet als *bomb*. Es lassen sich zusätzlich auch ähnliche Dokumente zu einem Treffer anzeigen, wobei die am höchsten gewichteten Begriffe, die im Dokument durch die

Abbildung 6: Schlagzeilen-Ansicht Ergebnisliste.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

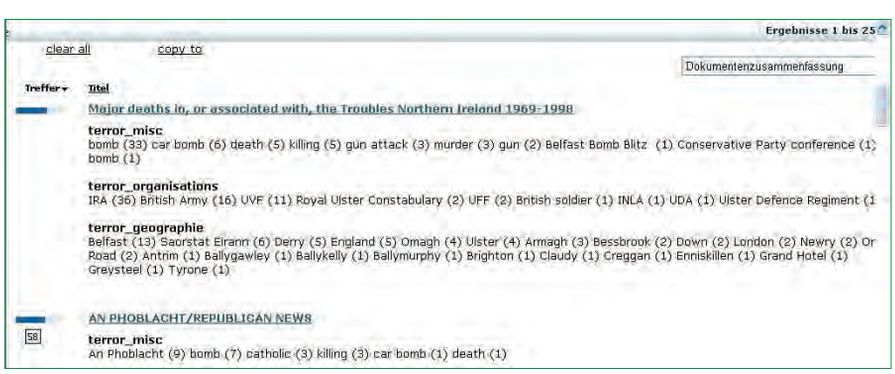

Abbildung 7: Ansicht Dokumentzusammenfassung Ergebnisliste.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

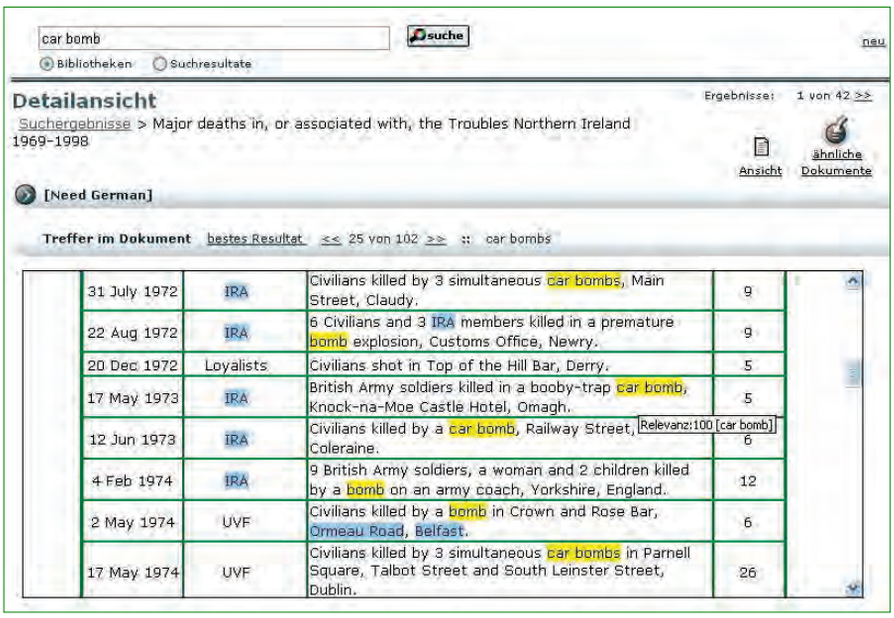

Abbildung 8: Detailansicht eines Dokuments aus der Ergebnisliste.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

initiale Suche gefunden worden sind, verwendet werden.

Sprachübergreifendes Retrieval

Der hier behandelte Thesaurus „Terror in Nordirland" wurde monolingual in Englisch erstellt. Convera bietet aber als Standard die Übersetzungen in die Sprachen Deutsch, Englisch, Französisch, Italienisch und Niederländisch an. So führt auch eine Suche nach dem Begriff *Hungerstreik*, wenn die Abfragesprache „Deutsch" gewählt wird, zu Dokumenten mit dem Begriff *hunger strike*, wie auch zu Dokumenten mit der Übersetzung in den erwähnten anderen Sprachen (Abbildung 9). Um eine sprachübergreifende Konzeptsuche zu ermöglichen, müsste der Thesaurus allerdings multilingual erstellt und gepflegt werden.

In der Suchoberfläche von RetrievalWare besteht die Möglichkeit, die Suchanfrage nicht nur um Terme der Konzepte in der Abfragesprache zu erweitern, sondern auch in anderen Sprachen. Dabei handelt es sich nicht um eine einfache Übersetzung der Suchterme, sondern um die Übersetzung in ein anderssprachiges Konzept. Markiert man die Option „Übereinstimmung mit der Abfragesprache", verhindert man dabei, dass Homonyme in anderen Sinnzusammenhängen gefunden werden, als in der Suchanfrage intendiert. Wörter mit gleicher Schreibweise in unterschiedlichen Sprachen, aber mit verschiedenen Bedeutungen werden dann in Bezug auf das Konzept der Abfragesprache erweitert. Da ein Wort in verschiedenen Sprachen in einem unterschiedlichen Sinnzusammenhang stehen kann, ist hier mit großer Sorgfalt vorzugehen.

Der Übersetzer muss über detaillierte Kenntnisse der Zielsprache verfügen, um das Konzept eines Begriffes in seinem nationalen, kulturellen Zusammenhang zu übertragen. Dem Aspekt des Sprachwandels kommt dabei eine besondere Bedeutung zu, weil sich die Bedeutung eines Wortes verändern kann. Nicht zu vergessen sind dialektale Besonderheiten oder unterschiedliche Wortbedeutungen innerhalb sozialer Gruppen. Im Zusammenhang mit Themen, die den Terrorismus oder die organisierte Kriminalität betreffen, müsste zusätzlich Insiderwissen über den Sprachgebrauch unter den Mitgliedern der entsprechenden Gruppierungen vorhanden sein, bis hin zu steganografischen Verfahren.

Erweiterte Suche

In der erweiterten Suchoberfläche sind vielfältige Einstellungen möglich, um die Suche zu optimieren und so den Recall, also die Vollständigkeit der Suchergebnisse, und die Precision, das heißt die Genauigkeit der Suchergebnisse, zu erhöhen. Es werden vier Suchtypen angeboten, die Boolesche Suche, eine feldspezifische Suche, Muster-Suche und Konzept-Suche (Abbildung 10).

(1.) Die Boolesche Suche erfolgt über die üblichen booleschen Operatoren ohne Relevance Ranking. (2.) Die Suchfelder der feldspezifischen Suche können in der administrativen Oberfläche passend zum Datenbestand erstellt werden. Bei Webseiten bieten sich z.B. Tags wie Description, Head und Body an. Das Feld „Dokumentinhalt muss enthalten" dient als Filter für eine Vorauswahl der Dokumente, auf welche dann die Suchanfragen aus dem primären Suchfeld und den spezifischen Feldern ausgeführt wird. Wird dort zum Beispiel *Belfast OR London* eingegeben und im primären Feld *car bomb*, so werden nur die Dokumente nach *car bomb* durchsucht, die entweder die Wörter *Belfast* oder *London* enthalten. (3.) In der Konzeptsuche wird je nach Einstellung das gesamte semantische Netz, das heißt das Konzept um den gesuchten Begriff, zum jeweiligen Anfrageterm hinzugezogen. Dabei kann der Konzepter-

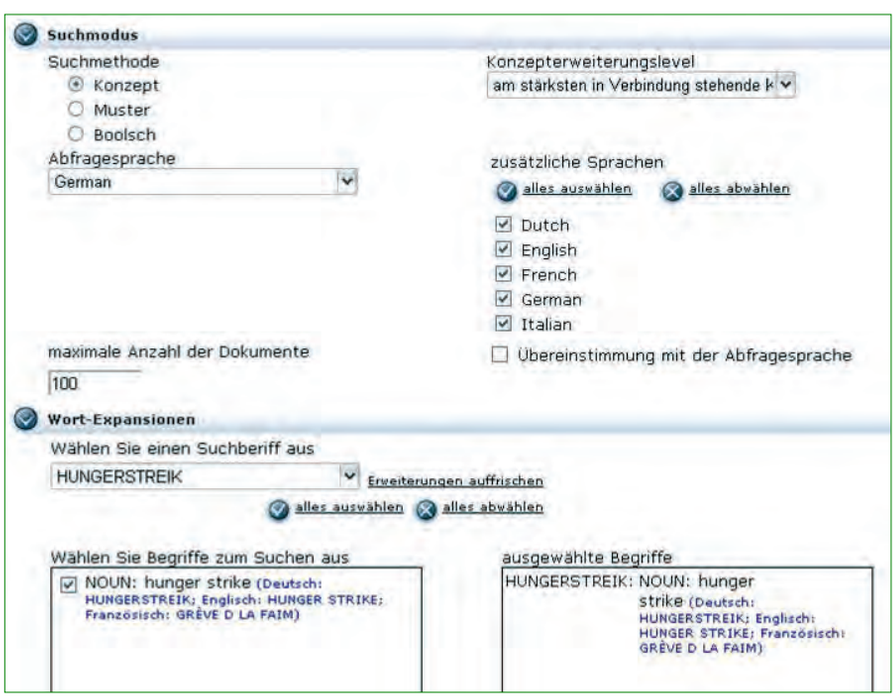

Abbildung 9: Cross-lingual Search in der erweiterten Suche.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

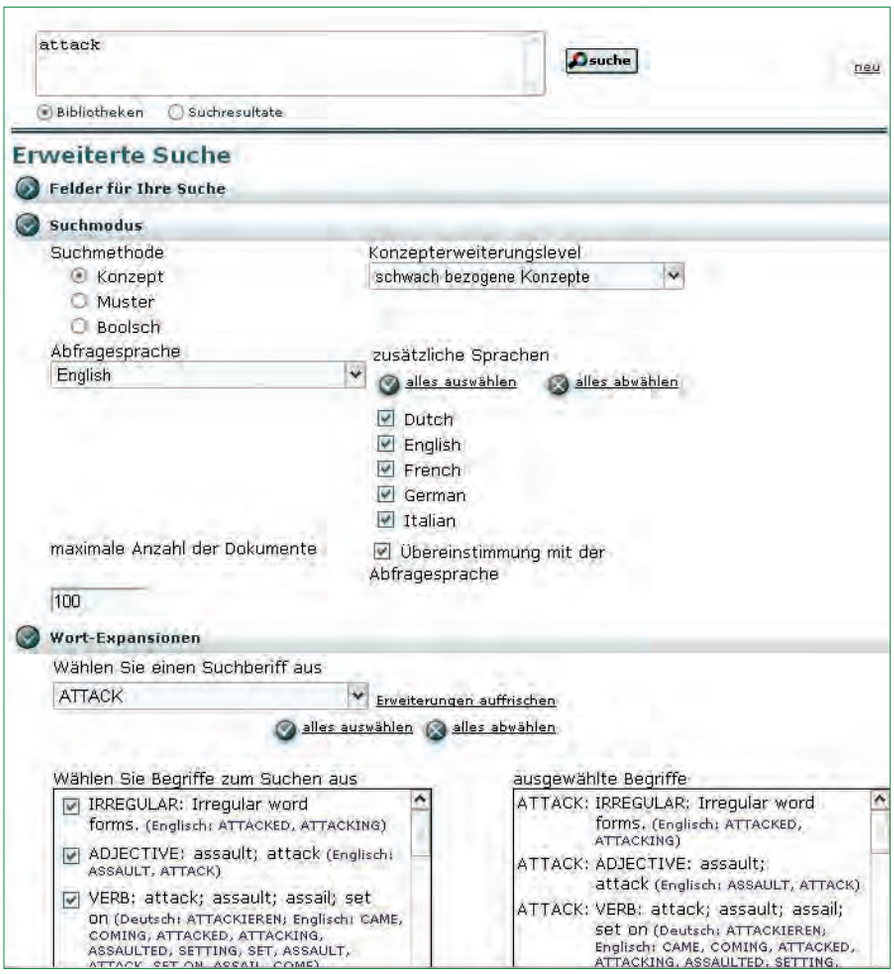

Abbildung 10: Erweiterte Suche.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

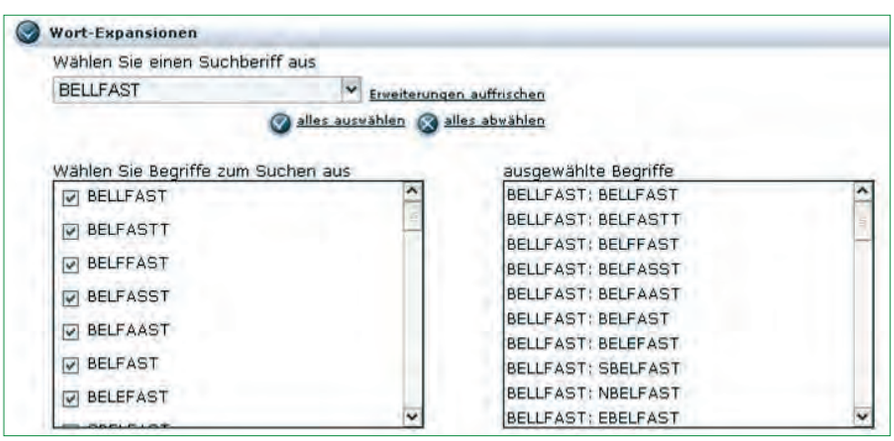

Abbildung 11: Wortexpansion der Mustersuche.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

weiterungsrahmen ebenfalls eingestellt und so die Granularität des semantischen Netzes justiert werden, um das der Anfrageterm erweitert wird. Zur Wahl stehen: exakte Suche, einfache Variationen, am stärksten in Verbindung stehende Konzepte, stark in Verbindung stehende Konzepte und schwach bezogene Konzepte. Je weiter der Expansionsrahmen gefasst wird, desto mehr Suchergebnisse werden angezeigt, da mehr Begriffe in die Suche und Gewichtung einbezogen werden. In der Wortexpansion werden zu den Termen konzeptuelle Erweiterungen angeboten, die man markieren und in die Suche einbeziehen kann. Die ursprüngliche Suchanfrage wird mit diesen Erweiterungen aufgefrischt. Dabei werden alle Konzepte der gewählten Sprache angezeigt. Eine Steigerung des Recalls geht bei schwach bezogenen Konzepten einher mit einer niedrigeren Precision. Hier muss der Nutzer entscheiden oder ausprobieren, ob seine Anfrage konkret oder weiter gefasst beantwortet werden soll. (4.) In der Muster-Suche werden unterschiedliche Schreibweisen, Tippfehler oder Ungenauigkeiten in der Darstellung z.B. gescannter Dokumente berücksichtigt. Convera nutzt die adaptive Mustererkennung APRP (adaptive pattern recognition process), um diese Unregelmäßigkeiten zu kompensieren. Somit würde z.B. der libysche Staatschef Muammer al-Gaddafi auch unter den Schreibweisen Mu`ammar al-Qadhafi, Muammar Qaddafi, Mo'ammar Gadhafi, Muammar Kaddafi, Muammar Qadhafi, Moammar El Kadhafi, Mu'ammar Al Qathafi, Mohammer Q'udafi oder Mu'ammar al-Qaddafi sowohl als Suchbegriff, als auch in den Dokumenten erkannt und in den Suchergebnissen enthalten sein (Bachmann 2004). Sucht man in der IRA-Datenbank nach Dokumenten, die den Bezirk Belfast behandeln, und gibt stattdessen in der Konzeptsuche *Bellfast* ein, so wird als Ergebnis ein einziges Dokument angezeigt, welches zufällig den gleichen Schreibfehler enthält. Die Mustersuche ermöglicht eine Auswahl möglicher Schreibweisen in der Wortexpansion (Abbildung 11). Die Ergebnisliste enthält daraufhin Dokumente, in denen der Begriff *Bellfast* in der korrekten Schreibweise „Belfast" vorkommt, sowie dessen verwandte Begriffe (Abbildung 12).

Es ist möglich, mit Hilfe von Operatoren die Suchtypen zu kombinieren. So kann man in der Muster- und Booleschen Suche auch die Konzeptsuche integrieren, indem man dem Term ein Ausrufungszeichen anfügt. Dieser Term wird dann semantisch um seine verwandten Begriffe erweitert, so dass z.B. Tippfehler des Nutzers oder unterschiedliche Benennungen von Begriffen zugleich mit einer Konzepterweiterung des vom System korrigierten Terms kombiniert werden. Die feldspezifische Suche ist kombinierbar mit allen drei Suchtypen. Die Dokumente müssen dann alle Terme der Felder und, optional des primären Suchfeldes, aufweisen, um in der Ergebnisliste angezeigt zu werden. Die Mustersuche kann ebenso in der Konzept- und Booleschen Suche integriert werden, indem man dem Suchterm eine Tilde ~ voranstellt. Anführungszeichen, das heißt Phrasen, verhindern jeweils die semantische Begriffserweiterung.

Fazit

Die zielgenaue Recherche in sehr großen Datenmengen – wie beim *Dark Web* – erfordert den Einsatz von hocheffizienten Mechanismen, die zum einen die relevanten Webseiten identifizieren und es zum anderen ermöglichen, die so ge-

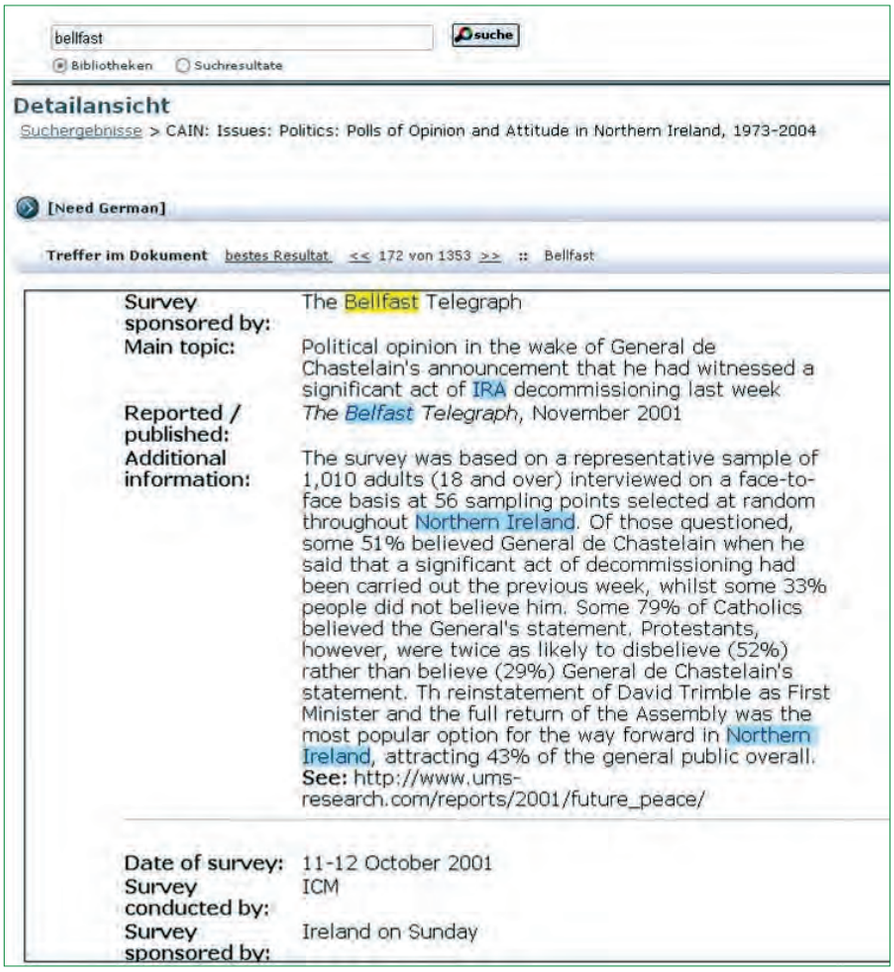

Abbildung 12: Detailansicht eines Ergebnisses aus der Mustersuche mit dem Term „Bellfast“. Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

wonnenen Dokumente durchsuchbar und darin enthaltene Informationen auffindbar zu machen.

Am Beispiel des Thesaurus Terror Nordirland konnte exemplarisch veranschaulicht werden, wie die Integration facettierter Thesauri und dynamischer Klassierung zu einer optimalen Auswertung auch sehr großer Datenmengen führt.

Durch die dynamische Klassierung wird die Datenmenge, auf der eine Suchanfrage ausgeführt wird, automatisch klein gehalten. Die vielen irrelevanten Dokumente, die auch bei einem fokussierten Crawling gesammelt werden, treten gar nicht in Erscheinung. Bei jedem Suchvorgang, ob beim Browsen durch die Verzeichnisse oder durch Eingabe eines Suchbegriffs, werden die passenden Ergebnisse in die Kategorien eingeordnet. Die Suche basiert dadurch auf einer bereits automatisch ermittelten Ergebnismenge. Durch den Einsatz facettierter Thesauri wird diese Ergebnismenge wiederum auf bestimmte Aspekte eingegrenzt. Facettierte Thesauri sind schnell einsetzbar und können nach und nach modifiziert werden. Die intellektuelle Thesaurus-Erstellung und -Pflege ermöglicht es, die einzelnen Thesauri bei Veränderungen im Sprachgebrauch oder Wissenszuwachs innerhalb kurzer Zeit anzupassen. Automatische Verfahren können ergänzend hinzugezogen werden. Der Aufbau eines semantischen Netzes erfordert aber auch bei automatischer Wortgewinnung eine intellektuelle Bearbeitung der Wissensordnung.

Mit Hilfe der dynamischen Klassierung und dem Einsatz facettierter Thesauri ist ein multidimensionales Retrieval möglich, das ein Höchstmaß an Flexibilität während der Suche bietet. Es werden nicht nur schnell präzise Ergebnisse gefunden, es ist auch eine ungezielte Recherche möglich, durch die assoziativ neue Erkenntnisse gewonnen werden. Der Nutzer kann zwischen mehreren thematischen Perspektiven wählen oder diese kombinieren und davon ausgehend auf Spezifikationen fokussieren, die ebenfalls miteinander kombinierbar sind. Insbesondere die Tabellenansicht ermöglicht eine assoziative Suche, die Schwerpunkte, Muster und Abweichungen sichtbar macht. Der Einsatz von dynamischer Klassierung und facettierter Thesauri bildet die optimale Grundlage für präzise, ballastarme Informationsgewinnung, insbesondere bei sehr großen Datenmengen.

Besonders im Fall polizeilicher Ermittlungen im Umfeld des *Dark Web* sollte jede Möglichkeit ausgeschöpft werden, um das Ergebnis so präzise wie möglich zu gestalten und zusätzlich eine nutzerfreundliche, intuitive Suche zu ermöglichen. Die Analyse von *Dark Web*-Inhalten kann nicht bei der Identifizierung von Verbindungen und technischen Möglichkeiten verharren. Ermittlungsmaßnahmen erfordern eine genauere Auseinandersetzung mit dem zusammenzustellenden Material und den darin enthaltenen Informationen. Der Zeit- und Kostenfaktor für den intellektuellen Ontologieaufbau wird durch den Zeitgewinn und die Effizienz bei der Informationsgewinnung wieder aufgefangen.

Danksagung

Wir danken Sonja Gust von Loh für die Bereitstellung von Informationen über „Check the Web“ und für Anregungen zum derzeitigen Stand der Dark-Web-Forschungen. Ebenfalls danken wir Jasmin Schmitz für die Überprüfung der Übersetzung des Abstracts ins Englische.

Literatur

Aitchison, J., *Gomershall*, A., & *Ireland*, R. (1969). Thesaurofacet: A Thesaurus and Faceted Classification for Engineering and Related Subjects. Whetstone, Leicester: English Electric Co. Ltd.

Artificial Intelligence Lab (2007). University of Arizona. Management Information Systems (MIS) Department. http://ai.bpa.arizona.edu/research/terror/index.htm [15.01.2008].

Bachmann, H. (2004). Convera Overview. Workshop on Managing Nuclear Knowledge. International Atomic Energy Agency. Trieste. http://www.iaea.org/inisnkm/nkm/documents/trieste2004/ L12(Bachmann).pdf [15.01.2008].

Bayer, O. et al.(2005). Evaluation of an ontology-based knowledge management system. A case study of Convera RetrievalWare 8.0. Information Services & Use 25(2005), 181-195.

Bergman, M.K. (2001). The Deep Web: Surfacing hidden value. The Journal of Electronic Publishing, 7(1). www.press.umich.edu/jep/07-01/bergman.html [15.01.2008].

Broughton, V. (2006). The need for a faceted classification as the basis of all methods of information retrieval. Aslib Proceedings 58(2006)1/2, 49-72.

Buchanan, B. (1989). Bibliothekarische Klassifikationstheorie. München: Saur.

Bundesministerium des Innern (2007). Europa sicher leben. www.eu2007.bmi.bund.de/ nn_1035938/EU2007/DE/Bilanz/GesamtBilanz__Ratspraesidentschaft__de.html [15.01.2008].

Chakrabarti, S., *van den Berg*, M., & *Dom*, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. Computer Networks 31(1999), 1623-1640.

Chen, H. et al. (2003). COPLINK* Managing Law Enforcement Data and Knowledge. Communications of the ACM 46(2003)1, 28-34.

Chen, H. et al. (2005). The dark web portal: Collecting and analyzing the presence of domestic and international terrorists groups on the web. Lecture Notes in Computer Science 3495, 623-624.

Cho, J. & *Garcia-Molina*, H. (2000). The evolution of the web and implications for an incremental crawler. In: Proceedings of the 26th International

Conference on Very Large Databases (S. 200-209).

Davison, B. (2000). Topical locality in the Web. In: Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval (S. 272–279). New York: ACM.

Demchak, C. C., *Friis*, C. & *La Porte*, T. M. (2000). Webbing Governance: National Differences in Constructing the Face of Public Organizations. In: *Garson*, G. D. (Ed.), Handbook of Public Information Systems (S. 71-84). New York: Marcel Dekker Publishers.

Elliott, A. (2001). Flamenco Image Browser: Using Metadata to Improve Image Search During Architectural Design. In: Proceedings of CHI 2001 (S. 69-70). New York: ACM.

English, J., *Hearst*, M., *Sinha*, R., *Swearingen* K. & *Yee*, P. (2002a). Hierarchical Faceted Metadata in Site Search Interfaces. CHI 2002 Conference Companion. http://flamenco.berkeley.edu/papers/chi02_short_paper.pdf [15.01.2008].

English, J., *Hearst*, M., *Sinha*, R., *Swearingen*, K. & *Yee*, P. (2002b). Flexible search and browsing using faceted metadata. Unpublished Manuscript. http://flamenco.berkeley.edu/papers/flamenco02.pdf [15.01.2008].

EU Council Secretariat (2007). The European Union and the Fight Against Terrorism. Brussels: Press Office of the Council of the European Union.

Hearst, M., *Elliott*, A., *English*, J., *Sinha*, R., *Swearingen*, K. & *Yee*, K.P. (2002). Finding the flow in web site search. Communications of the ACM 45(2002)9, 42-49.

Lewandowski, D. (2005). Web Information Retrieval: Technologien zur Informationssuche im Internet. Frankfurt am Main: DGI.

Kwasnik, B.H. (1999). The role of classification in knowledge representation and discovery. Library Trends 48(1999)1, 22-47.

Kwiatkowski, M., & *Höhfeld*, S. (2007). Thematisches Aufspüren von Web-Dokumenten – Eine kritische Betrachtung von Focused Crawling-Strategien. Information - Wissenschaft und Praxis, 58(2007)2, 69-82.

Li, J., *Zheng* R. & *Chen*, H. (2006). From fingerprint to writeprint. Communications of the ACM 49(2006), 76-82.

Netcraft Web Server Survey (2007). http://news.netcraft.com/archives/web_server_survey.html [15.01.2008].

Pant, G., *Srinivasan*, P. & *Menczer*, F. (2004). Crawling the Web. In: *Levene*, M., & *Poulovassilis*, A. (Eds.), Web Dynamics (S. 153-178). Heidelberg: Springer.

Qin, J. et al. (2005). The Dark Web Portal project: Collecting and analyzing the presence of terrorist groups on the Web. Lecture Notes in Computer Science 3495, 623-624.

Qin, J. et al. (2007). Analyzing terror campaigns on the internet: Technical sophistication, content richness, and Web interactivity. International Journal of Human-Computer Studies 65(2007), 71-84.

Ranganathan, S.R. (1933): Colon Classification. Madras: Madras Library Association.

Reid, E. & *Chen*, H. (2006). Extremist Social Movements Groups and Their Online Digital Libraries. Information Outlook 10(2006)6, 57-65.

Reid, E. et al. (2005). Collecting and analyzing the presence of terrorists on the Web: A case study of Jihad Websites. Lecture Notes in Computer Science 3495, 402-411.

Sacco, G. M. (2006). Dynamic taxonomies and guided searches. Journal of the American Society for Information Science and Technology 57(2006)6, 777-780.

Sherman, C. & *Price*, G. (2001). The Invisible Web: Uncovering Information Sources Search Engines Can't See. Medford, NJ: Information Today

Stock, M. (2002). Factiva.com: Neuigkeiten auf der Spur. Searches, Tracks und News Pages bei Factiva. Password 17(2002)5, 31-40.

Stock, W.G. (2007). Information Retrieval. Informationen suchen und finden. München, Wien: Oldenbourg.

Stock, W.G. & *Stock*, M. (2008). Wissensrepräsentation. Informationen auswerten und bereitstellen. München, Wien: Oldenbourg.

Uddin, M.N. & *Janecek*, P. (2007). The implementation of faceted classification in web site searching and browsing. Online Information Review 31(2007)2, 218-233.

Vickery, B.C. (1969). Facettenklassifikation. München-Pullach, Berlin: Verlag Dokumentation.

Yee, K.P. et al. (2003). Faceted metadata for image searching and browsing. In: Proceedings of the SIGCHI conference on Human factors in computing systems (S. 401-408). New York: ACM.

Zhou Y. et al. (2005). U.S. domestic extremist groups on the Web: Link and content analysis. IEEE Intelligent Systems 20(2005)5, 44-51.

Wissensordnung, facettierte Wissensordnung, dynamisches Klassieren, Information Retrieval, Dark Web, IRA-Datenbank, Convera RetrievalWare, facettierte Klassifikation, facettierter Thesaurus, Thesauruserstellung, Crawling, sprachübergreifendes Retrieval, semantisches Netz

Die Autoren

Silke Heesemann

ist Mitarbeiterin der Arbeitsgruppe Datenbanken und Softwaretechnik der Abteilung Informationswissenschaft an der Heinrich-Heine-Universität Düsseldorf.

silke.heesemann@uni-duesseldorf.de

Hans-Dieter Nellißen

ist technischer Mitarbeiter der Arbeitsgruppe Datenbanken und Softwaretechnik der Abteilung Informationswissenschaft an der Heinrich-Heine-Universität Düsseldorf.

hans.nellissen@uni-duesseldorf.de