GRACE - Eine Grid-basierte und kategoriebildende Suchmaschine

von Frank Scholze

1. Einleitung

GRACE (Grid Search and Categorization Engine) ist ein Projekt der Europäischen Union im fünften Rahmenprogramm der *Information Society Technologies* (IST)[1] Projekte. Koordiniert und geleitet wird es von der Forschungseinrichtung der Italienischen Telekom (Telecom Italia Lab). Zusammen mit der israelischen Firma Virtual Self trägt sie auch einen Großteil der technischen Entwicklung. Die wissenschaftlichen Entwicklungspartner des Konsortiums sind die School of Computing and Management Sciences der Sheffield Hallam University und das Europäische Zentrum für Hochenergiephysik CERN (Centre Européenne pour la Recherche Nucléaire), das als einer der Initiatoren des European DataGrid Projects[2] vor allem seine Erfahrung im Bereich Grid-Technologie einbringt. Die informationstechnische und wissenschaftliche Anwenderseite wird durch die Universitätsbibliotheken Stockholm und Stuttgart vertreten. Das Projekt begann im September 2002 und ist zunächst mit einer Laufzeit von 30 Monaten ausgestattet.

GRACE zielt mit seinem Ansatz einer verteilten Suche auf einen wissenschaftlichen Nutzerkreis. Hierbei sollen nicht nur die bekannten Informationsressourcen des WWW (Websites, Datenbanken, Kataloge etc.) für die Suche zur Verfügung stehen, sondern auch Grid-basierte Datenquellen. Die Ergebnisse der Suche in den erwartungsgemäß stark heterogenen Quellen werden weder in einer einheitlichen Ergebnisliste noch in einer nach Quellen unterschiedenen Ergebnisanzeige dargestellt, sondern nach inhaltlichen Kriterien in Gruppen zusammengefasst (kategorisiert bzw. geclustert). Dublette Ergebnisse sollen hierbei nach Möglichkeit nur innerhalb einer Kategorie eliminiert werden. Die Bezeichnungen der Kategorien werden dabei ad hoc mit Hilfe linguistischer Methoden aus den jeweiligen Ergebnissen generiert.

Im Folgenden soll auf die innovativen Aspekte Kategoriebildung und Grid-Technologie, mit denen sich das Projekt auseinandersetzt, näher eingegangen werden.

2. Grid-Computing

Die Idee des Grid-Computing entstand Anfang der 90er Jahre des vergangenen Jahrhunderts. Sie ist mit der Vorstellung von Stromnetzen zu vergleichen. Strom wird heute in einer einheitlichen Form bereitgestellt, ohne dass sich der Endverbraucher um dessen Erzeugungsweise oder Verteilung kümmern muss. Er benötigt lediglich einen (zumindest national) standardisierten Anschluss für die Geräte, die er betreiben möchte. Der Zustand im Bereich Computernetzwerke ist heute der Stromversorgungssituation um 1910 vergleichbar. Jeder, der computerbasierte Dienstleistungen (damals elektrische Geräte) betreiben will, muss sich um die notwendigen Rechnerkapazitäten (damals Stromgeneratoren) selbst kümmern.[3] Es gibt bereits Millionen von Computern mit entsprechender Rechenleistung und Speicherkapazität, die über das Internet miteinander verbunden sind. Was fehlt, ist die Infrastruktur und Standardisierung, um dieses Potential einheitlich und transparent zu nutzen. Hier setzt das Konzept des Grid-Computing ein. Das einfachste Szenario für das Netzwerk der Zukunft besteht aus vier Schritten:

- Der Benutzer beschreibt sein Anliegen, seine Aufgabe oder Anfrage mit Hilfe einer graphischen Oberfläche und liefert eventuell benötigte Rohdaten.
- Die hierfür benötigten Ressourcen (Rechner- und Speicherkapazität etc.) werden automatisch berechnet, gesucht und zugewiesen.
- Die Bearbeitung der Aufgabe wird überwacht.
- Der Benutzer wird nach Erledigung der Aufgabe benachrichtigt und die Ergebnisse bereitgestellt.

Aufgrund der vom Benutzer spezifizierten Anforderungen werden derzeit funktional drei verschiedene

Arten von Grid-Netzwerken unterschieden: Compute Grids (Parallele Berechnung von Algorithmen), Data Grids (Berechnung großer Datenmengen) und Application Grids (Anwendungen).

Auf konzeptioneller Ebene entspricht die Vorstellung des Grid Computing der Virtuellen Organisation, die alle drei genannten Bereiche umfassen kann. In einer Virtuellen Organisation schließen sich Personen oder Institutionen mit ihren personellen und technischen Kapazitäten zusammen, um gemeinsame Aufgaben durch gemeinsame Nutzung aller Ressourcen zu lösen. Die britische eScience Initiative[4] ist ein Versuch in diese Richtung. Hier wurde ein dienstleistungsorientierter Ansatz gewählt, bei dem Nutzer und Anbieter von Grid-basierten Diensten bzw. deren Software-Agenten über die jeweiligen Nutzungsbedingungen verhandeln. Um diese Verhandlungen führen zu können, müssen standardisierte Informationen über Anforderungen und zur Verfügung stehende Ressourcen vorliegen. Diese Standardisierungen sollen unter anderem im Rahmen der Semantic Grid Initiative[5] erfolgen.

GRACE versucht bereits zu einem sehr frühen Zeitpunkt an den Standardisierungen im Bereich des Semantic Grid mitzuwirken. Hierzu wird das Testbed des Europen DataGrid Projects benutzt. In diesem frühen Stadium liegen noch keine nennenswerten textuellen Datenbestände innerhalb des Grid vor. Projektziel ist daher auch die Einbindung einiger Volltextsammlungen - wie z.B. die Dokumentserver des CERN (CDS)[6] oder der Universität Stuttgart (OPUS)[7] - als sog. Storage Elements in eine Grid-Umgebung (vgl. *Abb.1*). Nur auf dieser Anwendungsebene können die Möglichkeiten der Grid-Technologie einem größeren Nutzerkreis zugänglich gemacht werden - würde sie fehlen, bliebe der Anwenderkreis auf wenige wissenschaftliche (Teil-) Gemeinschaften wie die Hochenergiephysik beschränkt, die große Datenmengen berechnen müssen.

Um eine sinnvolle Kommunikation der einzelnen Komponenten gewährleisten zu können, muss eine entsprechende Kommunikations- und Grid-Middleware ausgewählt werden. Dies könnte beispielsweise JXTA[8] sein. JXTA ist ein Bündel von Protokollen, um peer-to-peer-Netzwerke aufbauen zu können, das auch von einer Reihe weiterer anwendungsorientierter Grid-Projekte untersucht und voraussichtlich genutzt wird.[9] Als Grid-Middleware wird das im European DataGrid Project verwendete Globus toolkit eingesetzt, das sich derzeit im Übergang auf die Open Grid Services Architecture (OGSA)[10] befindet.

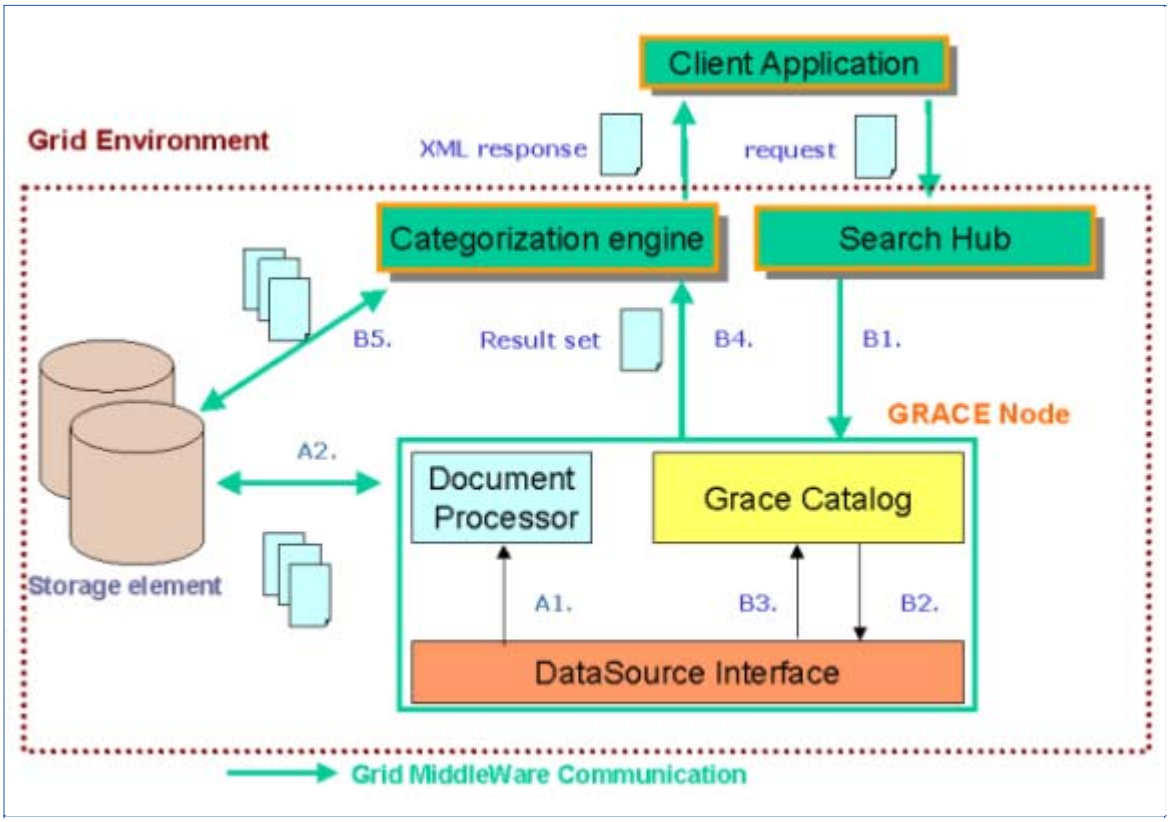

Abb.1: GRACE Architektur innerhalb einer Grid-Umgebung

3. Kategoriebildung

Document Clustering oder die sogenannte Bündelung von Dokumenten ist nichts anderes als eine Zuordnung von Dokumenten zu automatisch aus ihnen gewonnenen inhaltsrelevanten Stichworten oder

Phrasen (im Folgenden auch als Klassen oder Kategorien bezeichnet), d.h. eine Form automatisierter Inhaltserschließung. Hierzu existieren verschiedene meist graphenbasierte statistische Verfahren.[11] Produktiv im Bereich Information Retrieval bzw. für Suchmaschinen werden bislang nur wenige eingesetzt. Hierzu zählen u.a. der IBM Intelligent Miner for Text[12] und die Meta-Suchmaschine Vivisimo[13]. Bei Vivisimo wird ein conceptual clustering Algorithmus verwendet. Auch dies ist ein statistisches Verfahren, das jedoch nicht nur jeweils zwei Dokumente vergleicht, sondern den Gesamtkontext zu allen anderen Dokumenten berücksichtigt.[14] Hierbei ergeben sich zwei Problembereiche. Zum einen sollen möglichst alle gefundenen Dokumente einer Klasse oder Kategorie zugeordnet werden. Zum anderen stehen die Begriffe, welche die Kategorien beschreiben, in einem Spannungsverhältnis. Sie sollen einerseits nicht zu komplex sein, um Dokumente sicher Kategorien zuweisen und zwischen Kategorien unterscheiden zu können (Complexity). Andererseits sollen die Begriffe möglichst nur Dokumente beschreiben, die auch klassifiziert werden (Sparseness). Dies impliziert jedoch, dass die Begriffe eher komplexerer Art sind. Bestehende Algorithmen versuchen, eine Balance zwischen diesen Anforderungen zu finden.

Um die Dokumente linguistisch bearbeiten zu können, müssen sie in eine dafür geeignete Form gebracht werden. Dies geschieht bei GRACE ausschließlich durch die Entfernung von Stoppwörtern und eine morphologische Reduktion auf Grundformen (Document Processor, vgl. *Abb.1*).

Im Gegensatz zu den erwähnten Clustering-Verfahren erfolgt die Kategoriebildung in GRACE durch [15] Dies ist ein - wenn auch in einzelnen Sprachen unterschiedlich ausgeprägtes - sprachübergreifendes Konzept. Es geht davon aus, dass spezifische Sachverhalte gemäß sprachlichen Konventionen mit einer gewissen Varianz immer wieder gleich bezeichnet werden. Dies trifft in noch höherem Maß auf wissenschaftliche Texte zu. Als Beispiel können hier drei Texte zur Parkinsonschen Krankheit dienen.

Some drugs are known as dopamine agonists. These drugs bind to dopamine receptors in place of dopamine and directly stimulate those receptors. Some dopamine agonists are currently used to treat **Parkinson's disease**
dopamine neurons.

Quelle: http://www.utexas.edu/research/asrec/dopamine.html

The major goal of our laboratory is to identify the processes underlying the degeneration of neurons in **Parkinson's disease** and related disorders. **Parkinson's disease** is a chronic neurological disease which is characterized pathologically primarily by the loss of **dopamine neurons** of the **substantia nigra pars compacta**.

Quelle: http://cpmcnet.columbia.edu/dept/neurology/ni/research/Burke.html

Our laboratory is investigating the physiological mechanisms through which **dopamine**-containing **neurons** in the ventral midbrain encode information. Located in the **pars compacta** of the **substantia nigra** and in the adjacent ventral tegmental area, these cells form reciprocal connections with neurons in the basal ganglia, cortex and limbic forebrain. **Dopamine neurons** and their circuits are critically involved in the regulation of motor activity and in the motivational processes underlying learning and execution of goal-directed behaviors.

Quelle: http://www.mprc.umaryland.edu/faculty/shepard.html

Obwohl sie von unterschiedlichen Autoren stammen, gehorchen die Texte doch den lexikalischen Konventionen, indem immer wiederkehrende Ausdrücke benutzt werden (hier **fett** gekennzeichnet). Durch diese Konventionen ist es möglich, Bedeutungen in einem automatisierten Verfahren abzuleiten, auch ohne die Bedeutung jedes einzelnen Wortes zu verstehen.

Die in den idiomatischen Ausdrücken auftretenden Mehrdeutigkeiten werden nach Möglichkeit durch Kontextbetrachtungen aufgelöst, d.h. es werden die Begriffe betrachtet, die in einem weiteren syntaktischen Zusammenhang mit dem zu extrahierenden Begriff stehen (sog. Kollokationen). Klassisches Beispiel im Deutschen ist der Begriff "Bank". Steht er häufiger in Verbindung mit Begriffen wie Deutsche, Geld, Zinsen oder Beratung, ist auf eine andere Bedeutung zu schließen als in Kollokation mit Begriffen wie sitzen, liegen oder Park. Dies kann zu einer Erweiterung des Begriffs oder des Ausdrucks führen, der zur Kategoriebildung verwendet wird. Im vorliegenden Beispiel würden Kategorien für "Deutsche Bank" oder "auf der Bank sitzen" gebildet. Der Ausdruck "Bank im Park" würde auf diese Weise ebenfalls disambiguiert, wenn unterschiedliche Kollokationen vorliegen ("Geld

bei der Bank im Park anlegen" vs. "auf der Bank im Park sitzen").

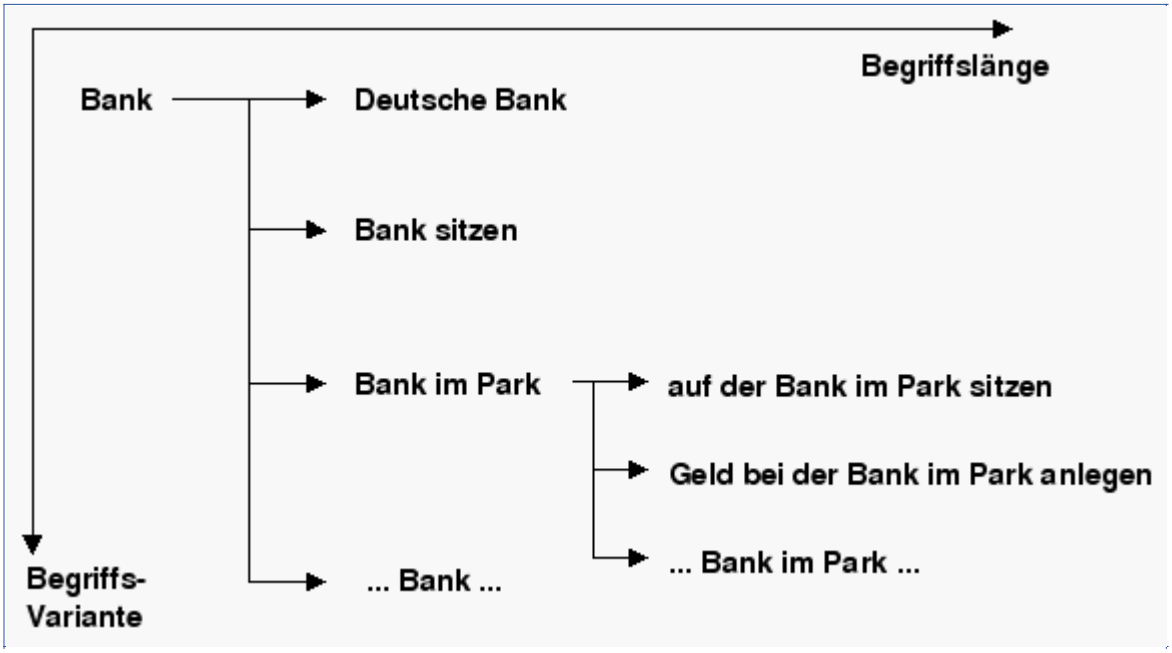

Abb. 2: Faktoren der Kontextbetrachtung

Abbildung 1 verdeutlicht das Gesagte. Bei stärkerer Kontextualisierung, d.h. zunehmender Länge des Begriffs, nimmt auch dessen Eindeutigkeit zu. Jede der Varianten kann auf diese Weise bis zu einem gewissen Grad disambiguiert werden.

Die extrahierten Begriffe werden als Knoten in nicht hierarchischen Concept Maps gespeichert. Die Verbindungen repräsentieren dabei die Häufigkeit des gemeinsamen Auftretens der Ausdrücke in den Dokumenten (Categorization Engine, vgl. *Abb.1*). Als Erweiterung soll im Rahmen des Projektes eine Kategoriebildung anhand vorgegebener Klassifikationen bzw. Thesauri realisiert werden. Sofern man über eine reine Suche nach Ähnlichkeitsbeziehungen zwischen Klassifikations- oder Thesaurusbegriffen und Dokumentbegriffen hinaus gehen will, besteht das Hauptproblem in der Abbildung der jeweils ad hoc extrahierten idiomatischen Ausdrücke auf die vorgegebenen Thesaurusbegriffe.

4. Ein Vergleich

Anhand des Prototypen eines Kategorisierungsalgorithmus, der im Projekt GRACE weiterentwickelt werden soll, wurde ein punktueller Vergleich mit der Metasuchmaschine Vivisimo durchgeführt, der sich ausschließlich auf die Bildung von Kategorien bezog. Als Quelle wurde jeweils die Suchmaschine AltaVista ausgewählt und mit der Phrase "document clustering techniques" gesucht. Die zu verarbeitende Treffermenge wurde auf 100 Dokumente begrenzt. AltaVista fand 40 Treffer, die jeweils auch verarbeitet wurden. Auffällig ist, dass die Komplexität der Begriffe bei Vivisimo heterogener war als bei GRACE. Begriffen mit geringer Komplexität (K=1) (z.B. Publications, Model oder Conference) standen hochkomplexe Ausdrücke wie "Comparison, Paper Presents The Results Of An Experimental Study" (K=9) gegenüber. Bei GRACE lag die Spanne zwischen K=5 und K=2, wobei 11 von 13 Begriffen mit K=2 eine sehr homogene Komplexität aufwiesen. Der komplexeste Ausdruck "Comparison of Document Clustering Techniques" war in Bezug auf die Anfrage auch der aussagefähigste. Die Zahl der nicht kategorisierten Dokumente war bei Vivisimo mit 13 gegenüber 7 deutlich höher. Der Preis für dieses bessere Ergebnis war die Laufzeit des Algorithmus bei GRACE. Da es sich jedoch bei GRACE noch um einen Prototypen handelt, der vermutlich auch unter anderen Hardwarebedingungen lief als Vivisimo, ist im Verlauf des Projektes mit einer Steigerung zu rechnen.

Suche nach "document clustering techniques"	**Vivisimo**	**GRACE**
Verarbeitete Dokumente	39	40
Nicht kategorisierte Dokumente	13	7
Zahl der Kategorien	14	13

Dauer der Kategorisierung	4 Sek.	40 Sek.
Gebildete Kategorien (mit zugeordneten Dokumenten)	Research (5) Automatically (2) Databases (2) Other Topics (1) Comparison, Paper Presents The Results Of An Experimental Study (3) Algorithm (3) Publications (3) Model (4) Conference (3) Template (2) Search Engines (2) Digital Library (2) Course, Computing (2) Other Topics (13)	Information Retrieval (13) Clustering Techniques (12) Document Clustering (11) Document Clustering Techniques (11) Clustering Web (5) Comparison of Document Clustering Techniques (3) Knowledge Technology (3) Search Engines (3) Searching Web (3) Computer Science (2) Search Tools (2) Subject Indexes (2) Other Documents (7)

Tab. 1: Vergleich der Kategorisierung bei Phrasensuche

In einem zweiten Vergleich wurde eine Suche mit der Booleschen Verknüpfung "und" durchgeführt. Die Suchbegriffe blieben unverändert, ebenso die Begrenzung der Verarbeitung auf 100 Dokumente ("+document +clustering +techniques"). AltaVista fand 26.463 Dokumente, d.h. die Ergebnismenge war bezüglich ihrer Precision, d.h. der Anzahl der relevanten Treffer, wesentlich schlechter als im ersten Versuch. Auffällig war, dass die Verarbeitungsdauer in beiden Fällen gleich blieb. Bei größeren Dokumentmengen stieg jedoch auch die Laufzeit bei GRACE überproportional an (Vivisimo benötigte ca. 7 Sek. gegenüber 60 Sek. bei 500 verarbeiteten Dokumenten). Während die Zahl der Kategorien ebenso wie die Zahl der nicht kategorisierten Dokumente bei GRACE konstant blieb, wuchs sie bei Vivisimo stark an. Die Komplexität der Kategorien blieb bei Vivisimo sehr heterogen, wobei die gering komplexen Begriffe (K=1) durchweg wenig aussagekräftig waren (Paper, Writing, References etc.). Die Komplexität bei GRACE blieb sehr homogen (K=2 bei 12 von 13 gebildeten Begriffen), die Aussagefähigkeit der Begriffe ging jedoch aufgrund des heterogeneren gefundenen Materials (geringere Precision aufgrund höheren Recalls) etwas zurück. Sie war jedoch immer noch deutlich höher als bei Vivisimo.

Suche nach +document +clustering +techniques	**Vivisimo**	**GRACE**
Verarbeitete Dokumente	98	101
Nicht kategorisierte Dokumente	34	9
Zahl der Kategorien	31	13
Dauer der Kategorisierung	4 Sek.	39 Sek.
Gebildete Kategorien (mit zugeordneten Dokumenten)	Algorithms (7) Software (2) Other Topics (5) Bibliography (9) Classification (4) Introduction. People. Bibliography (2) Other Topics (3) Paper (6) Comparison of Document Clustering Techniques (2) Other Topics (4)	Clustering Using (25) Clustering Algorithms (24) Cluster Technique (21) Used Techniques (21) Cluster Based (20) Information Retrieval (15) Data Mining (13)

	Classification (7) Bibliographic Remarks. In Recent Years (2) Retrieval (2) Other Topics (3) Data Clustering (4) Data Mining (4) Search engines (3) Dynamic (5) Fuzzy Clustering (2) Other Topics (3) Document Retrieval (4) Department of Computer (3) Writing (2) References 1 (3) Association Rule (2) Graph (3) Tools (3) Map (3) Syntactic Clustering Of The Web. Andrei Z. Broder (2) Accelerators (2) Visualization Techniques (2) Other Topics (34)	Clustering Data (9) Cache clustering (3) Graph Clustering (3) Internet Search (3) Method for Detecting (3) Other Documents (9)

Tab. 2: Vergleich der Kategorisierung bei Boolescher Suche

Zusammenfassend lässt sich feststellen, dass die Suche mit GRACE in diesem frühen Projektstadium noch sehr lange dauert. Die gebildeten Kategorien sind jedoch homogener und aussagekräftiger als bei Vivisimo, mithin ist daher auch eine leistungsfähigere inhaltliche Zuordnung der einzelnen Dokumente zu verzeichnen. Da mit einer Verbesserung der Performanz im Verlauf des Projektes zu rechnen ist, wird GRACE voraussichtlich vor allem bei einer größeren Anzahl relevanter Ergebnisse zu einer leistungsstarken inhaltlich erschließenden Suchmaschine werden.

5. Ausblick

Grid-Computing stellt eine neuartige technische Entwicklung von Rechnernetzen und Informationsverarbeitung dar, deren Auswirkungen für die Arbeitsweise Digitaler Bibliotheken erst noch erforscht und erprobt werden müssen. Im Rahmen des Projektes GRACE soll hierzu ein Baustein beigetragen werden. Es ist das erste Projekt, das sich mit der verteilten, kategoriebildenden Suche unter Einbeziehung von Grid-Netzwerken beschäftigt, und könnte der Anstoß für weitere Arbeiten auf diesem Gebiet sein.

Zum Autor

Frank Scholze ist Leiter der Abteilung Digitale Bibliothek der

Universitätsbibliothek Stuttgart
Pfaffenwaldring 55
D-70550 Stuttgart
Tel. 0711-6854731
E-Mail: scholze@ub.uni-stuttgart.de

Anmerkungen

1. http://www.grace-ist.org

2. http://www.eu-datagrid.org

3. *The grid: blueprint for a new computing infrastructure*. Ed. by Ian Foster, Carl Kesselman. San Francisco, 1999. Ch. 2 "A computational Grid is a hardware and software infrastructure that provides

dependable, consistent, pervasive and inexpensive access to high-end computational capabilities"

4. http://esc.dl.ac.uk/InfoPortal/

5. David De Roure, Nicholas Jennings, Nigel Shadbolt: *A Future e-Science Infrastructure*. Report commissioned for EPSRC/DTI Core e-Science Programme 2001 http://www.semanticgrid.org/v1.9/semgrid.pdf

6. http://cds.cern.ch

7. http://elib.uni-stuttgart.de/opus/

8. http://www.jxta.org

9. http://www-unix.globus.org/cog/projects/jxta/

10. http://www.globus.org/ogsa/

11. Vgl. Gheorghe Muresan: *Using document clustering and language modelling in mediated information retrieval.* Aberdeen, Robert Gordon University Diss., 2002. *Classification, clustering and data analysis : recent advances and applications*. Ed by Krzysztof Jajuga. Berlin, 2002

12. http://www-3.ibm.com/software/data/iminer/fortext/download/factsheet.pdf

13. http://vivisimo.com Vgl. http://www.pcwelt.de/ratgeber/online/19986/4.html

14. Ryszard S. Michalski, "Conceptual Clustering: A Theoretical Foundation and a Method for Partitioning Data into Conjunctive Concepts". In: *Optimisation et classification automatique - Textes des exposes du Seminaire organise par l'Institute de Recherche d'Informatique et d'Automatique* (IRIA) octobre 1978 - juin 1979, E. Diday (editor). Le Chesnay, 1979, pp. 254-294.

15. Method for Automatic Extraction of Key phrases from Text. Israel Patent Office 09/670,994 (September 27, 2000) http://www.vself.com