

Graphische Datenanalyse für digitale Bibliotheken – Leistungs- und Funktionsumfang moderner Analyse- und Visualisierungsinstrumente

DIETER SCHWARTZ

Das WorldWideWeb stellt umfangreiche Datenmengen zur Verfügung. Für den Benutzer wird es zunehmend schwieriger, diese Datenmengen zu sichten, zu bewerten und die relevanten Daten herauszufiltern. Einen Lösungsansatz für diese Problemstellung bieten Visualisierungsinstrumente, mit deren Hilfe Rechercheergebnisse nicht mehr ausschließlich über textbasierte Dokumentenlisten, sondern über Symbole, Icons oder graphische Elemente dargestellt werden. Durch geeignete Visualisierungstechniken können Informationsstrukturen in großen Datenmengen aufgezeigt werden. Informationsvisualisierung ist damit ein Instrument, um Rechercheergebnisse in einer digitalen Bibliothek zu strukturieren und relevante Daten für den Benutzer leichter auffindbar zu machen.

1 Einleitung

Informationsvisualisierung kann als eigenständige Forschungsdisziplin mit unterschiedlichen Forschungsfeldern betrachtet werden. Zu diesen Forschungsfeldern zählen die Bereiche Information Retrieval, Hypertext und WWW, Digitale Bibliotheken und Mensch-Computer-Interaktion [8, S. 1]. Die Begriffe Knowledge Visualization und Information Visualization sind dabei zu unterscheiden; zudem ist eine Abgrenzung dieser Begriffe zu Scientific Visualization und Simulation Modelling vorzunehmen.¹ Betrachtet man die Visualisierungstechniken – also die graphischen Elemente, die für eine Visualisierung eingesetzt werden – können konventionelle und innovative Techniken unterschieden werden. Den konventionellen Techniken sind u. a. Kreis-, Balken- oder Liniendiagramme, aber auch Glyphenplots zuzuordnen. Im Rahmen der innovativen Techniken wird eine Reihe sehr unterschiedlicher graphischer Darstellungsformen oder Metaphern angeboten, z. B. Hyperbolic Tree, Perspective Wall oder Information Landscape.²

¹ Zur Unterscheidung dieser Bereiche vgl. u. a. [13, S. 2; 23].

² Zu den unterschiedlichen Darstellungsformen vgl. [5; 15; 18].

Bereits vor der Entwicklung der Datenverarbeitung wurden komplexe Sachverhalte und umfangreiche Datenmengen graphisch dargestellt. So hat beispielsweise der französische Ingenieur Charles Joseph Minard die Verluste der Napoleonischen Truppen im Jahre 1812 aufgezeichnet. Die Truppenstärke, die sich von anfänglich 420 000 auf 100 000 Mann reduzierte, die jeweiligen Tagestemperaturen und die Entfernung von der polnischen Grenze bis Moskau wurden in einer Graphik zusammengefasst.³ Das Werk „The Commercial and Political Atlas“ von William Playfair aus dem Jahre 1786 veranschaulicht mit Hilfe von graphischen Darstellungen den Import und Export des Mutterlandes England von und nach Ost-Indien.⁴ Dieser Atlas gilt als eines der ersten Werke, die statistische Graphen nutzten.

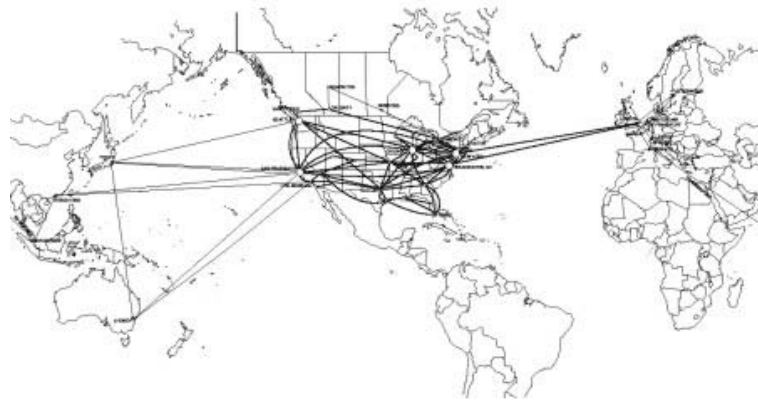


Abb. 1: Visualisierung von Datenströmen im globalen Datennetz

Diese historischen Beispiele veranschaulichen, dass in graphischen Darstellungen beschreibende Daten mit geographischen Positionen verbunden werden können. Ein aktuelles Beispiel für eine Visualisierung von geographischen und beschreibenden Daten liefert die Kartographie des Internets. Abbildung 1 zeigt den Datenverkehr in Nordamerika sowie die Verbindungen zum europäischen und pazifischen Raum.⁵

³ Abbildungen der Karte sind u. a. zu finden unter der URL:

http://en.wikipedia.org/wiki/Charles_Joseph_Minard.

⁴ Abbildungen sind u. a. zu finden unter der URL:

http://en.wikipedia.org/wiki/William_Playfair

⁵ Diese Graphik sowie weitere Darstellungen zum Datenverkehr sind zu finden bei [1].

Die Internet-Suchmaschine *mooter*⁸ führt eine Clusterung der Suchergebnisse durch und stellt Beziehungen zwischen den Clustern in Form von einfachen Netzen dar. Der Anwender findet mit Hilfe dieser Cluster übergeordnete Themenbereiche und kann auf deren Basis aus einer umfangreichen Treffermenge gezielt auswählen. Vom Netz kann er dann direkt zu den textbasierten Ergebnislisten gelangen (vgl. Abb. 3).

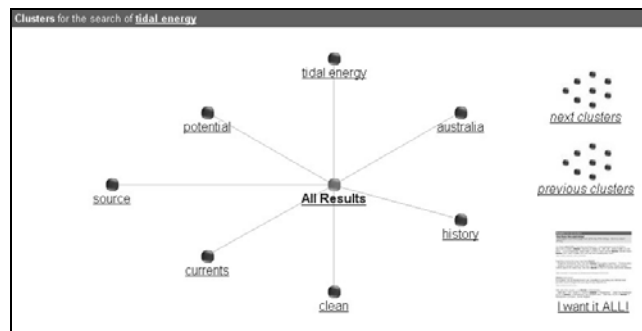


Abb. 3: Visualisierung mit Hilfe von *mooter*

Das GRIDL-System (**GR**aphical Interface for **D**igital Libraries) liefert eine Visualisierung für Suchergebnisse aus einer elektronischen Fachbibliothek.⁹

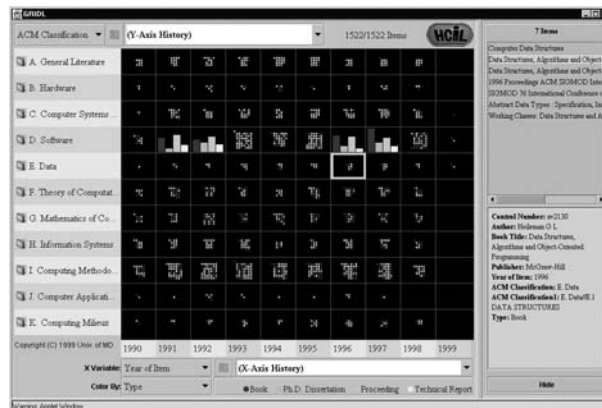


Abb. 4: Visualisierung mit Hilfe von *GRIDL*

⁸ Mooter-Web-Search. URL: <http://www.mooter.com>.

⁹ GRIDL, GRaphical Interface for Digital Libraries. URL: <http://www.cs.umd.edu/hcil/west-legal/gridl>.

Dieses System stellt Rechercheergebnisse in Form einer Matrix dar. Diese Matrix wird über die Systemstellen einer Klassifikation und die Erscheinungsjahre der Medien aufgespannt. Auf der GRIDL-Bildschirmoberfläche sind neben dieser Ergebnismatrix eine Liste aller markierter Titel und eine umfangreichere Titelaufnahme einer ausgewählten Publikation zu finden (vgl. Abb. 4).

Die Häufigkeit des Vorkommens von einzelnen Medientypen wird in Form von Histogrammen dargestellt. In den Histogrammen werden Bücher (rot), Dissertationen (grün), Proceedings (orange) und Technical Reports (gelb) entsprechend farblich hervorgehoben (vgl. Abb. 5).

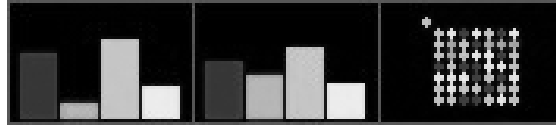


Abb. 5: Häufigkeit von einzelnen Medientypen

Die vorgestellten Software Tools spiegeln die unterschiedlichen Möglichkeiten wider, Rechercheergebnisse graphisch aufzubereiten. Die Systeme besitzen zudem vielfältige Features, um eine Interaktion mit den Daten durchzuführen. Ebenso wie serverbasierte oder datenbankspezifische Systeme ermöglichen lokale Visualisierungsinstrumente umfangreiche Interaktionsmöglichkeiten. Eine derartige lokale Anwendung ist BibTechMonTM.

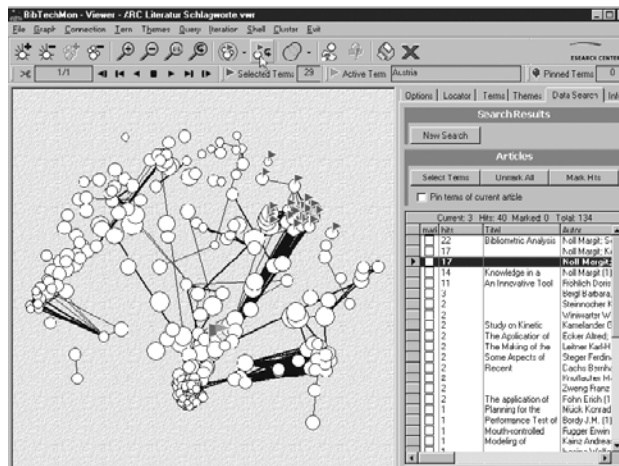


Abb. 6: Visualisierung mit Hilfe von BibTechMonTM

BibTechMon steht für **Bibliometrisches Technologie-Monitoring**.¹⁰ Das System eignet sich insbesondere dafür, Beziehungen zwischen Dokumenten in Form von Netzen darzustellen. Dokumente werden dabei als Knoten dargestellt, die Bedeutung eines Dokumentes wird durch die Größe des Knotens veranschaulicht. Beziehungen zwischen den Dokumenten werden durch Verbindungen (Kanten) zwischen den Knoten dargestellt, die Strichstärke dieser Verbindungen macht eine Aussage über die Stärke der inhaltlichen Beziehung zwischen den Dokumenten (vgl. Abb. 6). Das Modul Network Viewer von BibTechMonTM dient speziell der Objektanalyse.¹¹

3 Analyse von wissenschaftlicher Literatur

Nachfolgend werden die Vorgehensweise bei der Durchführung einer bibliometrischen Analyse sowie ein entsprechendes Software Tool vorgestellt. Beispielhaft wird jene wissenschaftliche Literatur ausgewertet, die bei der Entschlüsselung der DNS-Struktur von den Nobelpreisträgern J. Watson und M. Wilkins herangezogen wurde. Ein Netz aus zitierten und zitierenden Quellen, ein sogenanntes Zitatennetz, wird darauf aufbauend generiert.

3.1 Die Entschlüsselung der DNS-Struktur

Die Struktur der Desoxyribonukleinsäure¹² (DNS) wurde von James Watson und Francis Crick aufgeklärt, die im Jahre 1962 zusammen mit Maurice Wilkins den Nobelpreis für Medizin erhielten.¹³ Als weitere Personen, die einen wesentlichen Beitrag zur Entschlüsselung der DNS leisteten, sind Oswald Theodore Avery, Colin McLeod und Maclyn McCarty zu nennen. Diese konnten zeigen, dass die DNS Träger der Erbinformation ist. Für die

¹⁰ BibTechMon, URL: <http://www2.arcs.ac.at/S/ST/BTM/bibtech.engl.pdf>.

¹¹ Für die Berechnung des Netzes muss festgelegt werden, welche Objekte (z. B. Autorennamen, Schlagwörter) und welche Beziehungen zwischen diesen Objekten (z. B. Schlagwörter aufgrund der Häufigkeit ihres Vorkommens in den Dokumenten) visualisiert werden.

¹² Die Desoxyribonukleinsäure (DNS), meist nach der englischen Bezeichnung *deoxyribonucleic acid* mit DNA abgekürzt, ist ein Makromolekül, das bei der Vererbung als Träger der Information dient. Die Desoxyribonukleinsäure ist ein langkettiges Polymer, das aus vielen Einzelbausteinen besteht, die man Desoxyribonukleotide nennt.

¹³ Die DNS wurde allerdings bereits 1869 von Friedrich Miescher entdeckt, der in Zellkernen das Nuklein vorfand, jedoch die Funktion dieser Substanz noch nicht sicher bestimmen konnte.

Entschlüsselung der DNS-Struktur mussten entsprechende Analysetechniken vorhanden sein. In diesem Zusammenhang sind u. a. die Röntgenbeugungsdiagramme von Rosalind Franklin zu nennen. Die Zusammenarbeit von mehreren Wissenschaftlern in Teams und die Kooperation zwischen unterschiedlichen Forschungsdisziplinen sind Kennzeichen einer entstandenen Großforschung.¹⁴

3.2 Bibliometrische Analyse und Zitatanalyse

Mit Hilfe von bibliometrischen Analysen können Aussagen über die Wirkung von Veröffentlichungen gemacht werden, die von einzelnen Forschern, einer Forschergruppe oder einem Institut erstellt wurden. Es können damit Rückschlüsse auf die Publikationsleistung und Sichtbarkeit von Forschungsergebnissen gezogen werden [2, S. 15f.; 15; 16].

Die Zitatanalyse ist ein Teilgebiet der Bibliometrie; in ihr werden die Beziehungen zwischen Veröffentlichungen und den darin zitierten Quellen untersucht. Ausgangspunkt ist die Vorstellung, dass zitierte Quellen sichtbare Verknüpfung zwischen Veröffentlichungen sind. Zwei Zeitschriftenartikel werden beispielsweise als ähnlich erachtet, wenn sie auf ähnliche Quellen zurückgreifen.¹⁵ Grundlegende Arbeiten in diesem Bereich wurden u. a. von Garfield und Price geleistet [9]. Elektronische Systeme (z. B. Web of Science, Scopus oder IEL) bieten gegenwärtig komfortable Möglichkeiten zur Auswertung derartiger Beziehungen.¹⁶

¹⁴ Weiteres Kennzeichen ist der Aufwand an Finanz- und Personalmitteln, der die Großforschung (Big Science) zu einem wichtigen Element der nationalen Wirtschaft werden läßt. Zur Beziehung von Little Science und Big Science vgl. [24]. Zur Bedeutung des „information resources management“ in Bibliotheken und zur Rolle der digitalen Bibliotheken für die Big Science vgl. [22].

¹⁵ Anmerkung: Damit verbunden sind Probleme der Selbst-Zitation oder Negativ-Zitation.

¹⁶ Basis für das Web of Science ist der Science Citation Index und der Social Science Citation Index. Scopus ist eine Literaturdatenbank, die die weltweit größte Sammlung an Abstracts und Quellenverweisungen enthält. In der IEEE/IEE Electronic Library (IEL) werden unter „Abstract Plus“ die Rubriken References und Citing Documents angeboten,
URL: <http://www.thomsonisi.com>, <http://www.scopus.com>,
<http://www.ieee.org/ieeexplore>.

3.3 Analyse und ihre Instrumente

Die Software HistCite™ ist ein Instrument, mit dessen Hilfe besonders häufig zitierte Quellen auf der Grundlage der Daten des Science Citation Index identifiziert werden können.¹⁷ HistCite™ generiert Tabellen und Matrizen sowie Historiographen. Ausgangspunkt für diese Form der bibliometrischen Analyse ist die Vorstellung, dass die Geschichte der Wissenschaft als eine chronologische Abfolge von Ereignissen betrachtet werden kann, in der jede Entdeckung von früheren Entdeckungen abhängt. Um eine graphische Datenanalyse erzeugen zu können, müssen zunächst die Daten ausgewählt und in geordneter Form dargestellt werden.¹⁸

3.3.1 Vom Datensatz zur Zitatenmatrix

Ausgangspunkt für eine Analyse und die sich daran anschließende Visualisierung in Form eines Historiographen sind bibliographische Daten. In der nachfolgenden Abbildung 7 ist der Aufbau eines einzelnen Datensatzes aufgeführt. Als Autor erscheint der Nobelpreisträger M. Wilkins, der im Jahre 1953 in der Zeitschrift *Nature* diesen grundlegenden Artikel veröffentlicht hat. Die zitierten Quellen sind durch CR (cited references) gekennzeichnet.

Author(s)	WILKINS MHF; SEEDS WE; STOKES AR; WILSON HR
Title	HELICAL STRUCTURE OF CRYSTALLINE DEOXYRIBOSE NUCLEIC ACID
Journal	NATURE 172(4382):759-762
...	...
CR	ASTBURY WT, 1947, S SOC EXP BIOL, V1, P66 DONOHUE J, 1953, P NATL ACAD SCI USA, V39, P470 ... WATSON JD, 1953, NATURE, V171, P737 WILKINS MHF, 1953, NATURE, V171, P737 WILKINS MHF, 1951, NATURE, V167, P759 ZAMENHOF S, 1950, J BIOL CHEM, V187, P1

Abb. 7: Bibliographische Daten - Datensatz 38

¹⁷ HISTCITE™ – Bibliographic Analysis and Visualization Software, URL: <http://www.histcite.com/>.

¹⁸ Zum methodischen Problem bei der Auswahl der Dokumente, die für eine Auswertung herangezogen werden, vgl. [12].

In einer Matrix werden alle bibliographischen Daten zusammengefasst, zitierte Publikationen (cited nodes) werden zitierenden Publikationen (citing nodes) gegenübergestellt (vgl. Abb. 8). Ein Dokument wird in der Matrix als *node* bezeichnet, da es als Knoten innerhalb eines Zitatennetzes erscheint.

cited nodes	nodes	citing nodes
	1 1944 AVERY OT	4, 6 ... 100, 102 ... 331
	...	
33, 36	38 1953 WILKINS MHF	47, 71 ... 224, 230 ... 262
	...	
1, 2, 33, 34, 36, 38, 85	136 1955 FEUGHELMAN M	131, 142 ... 224, 225 ... 327
	...	

Abb. 8: Zitatennetz

Da eine Analyse häufig mehrere hundert Dokumente einschließt, wird die generierte Matrix entsprechend unübersichtlich. Eine Datenreduktion und eine einheitliche Strukturierung der Ergebnisse wird erforderlich und kann mit Hilfe der Datenverarbeitung erreicht werden.

3.3.2 Von der Matrix zum Zitatennetz

In Historiographen werden die Dokumente als Knoten dargestellt und die Beziehungen zwischen den Dokumenten durch Pfeile angezeigt. Die Dokumente sind in eine zeitliche Reihung gebracht, indem sie nach Publikationsjahren absteigend angeordnet sind. In Historiographen liegen auf einer horizontalen Ebene also Publikationen eines Jahrganges. Die zitierenden Quellen zeigen auf die zitierten Quellen. Die herausragende Stellung eines Dokumentes – und damit der Durchmesser des Knotens innerhalb des Netzes – ergibt sich durch die Anzahl der Verweisungen (vgl. Abb. 9). Für den Knoten 1 können sieben Pfeilspitzen, für den Knoten 33 ebenfalls sieben Pfeilspitzen gezählt werden. Die eigentliche Anzahl der Verweisungen ist jedoch ungleich höher, da durch geeignete Datenreduktion *überflüssige* Verweisungen ausgeblendet wurden.¹⁹

¹⁹ Vom Dokument 136 gehen 5 Pfeile aus, die sich dann weiter verzweigen. Das Dokument 136 verweist letztlich auf sieben Dokumente (1, 33, 43, 85, 36, 38, 2).

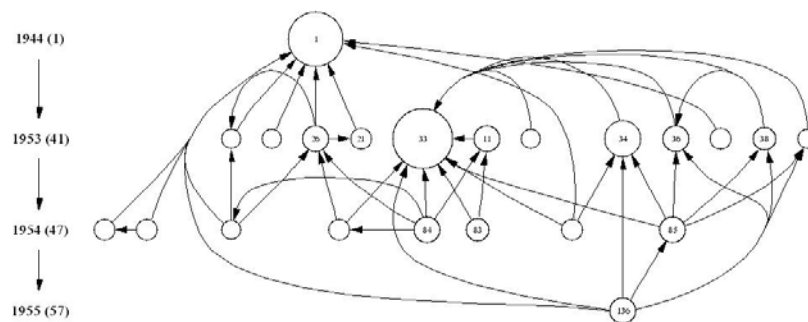


Abb. 9: Historiograph

Die in der Zitatenmatrix festgelegten Beziehungen wurden in den Historiographen übertragen, dabei wurden nicht alle Daten der Matrix in vollem Umfang umgesetzt. Durch Zusammenführen von mehreren Verweisungen zu einem Pfeil wurde die Gesamtzahl der Pfeile reduziert und damit eine vereinfachte, aber anschauliche Graphik erstellt. Ebenfalls zur Steigerung der Übersichtlichkeit wurden bestimmte Knoten nicht mehr nummeriert.²⁰

3.4 Zusammenhalt in Zitennetzen

Das chronologische Ausrollen von Publikationen und das Verbinden aufgrund der zitierten Quellen ist eine Methode, um eine Strukturierung von Dokumenten vorzunehmen. Stellt man nicht die zeitliche Abfolge, sondern die Stärke der inhaltlichen Beziehung in den Mittelpunkt der Betrachtung, so entsteht zwangsläufig ein anderer Graph. Über Gewichtungsfaktoren werden Hauptströme lokalisiert, dabei werden Verbindungen (Kanten) zwischen Dokumenten (Knoten) stark hervorgehoben.²¹ Auf der Grundlage von mathematischen Algorithmen können dann aus einem umfangreichen Gesamtsystem Subnetze generiert werden. Anwendung findet dieses Verfahren u. a. bei der Auswertung von Patentliteratur.²²

²⁰ Detaillierte Angaben zur Graphik sind im Folder Avery_Watson-Crick/ zu finden, URL:

<http://garfield.library.upenn.edu/histcomp/index-watson-crick.html>.

²¹ Vgl. die Darstellung über sogenannte Suchpfade (search path) bei [12].

²² Eine anschauliche graphische Datenanalyse für Patente findet man bei [3].

4 Analyse von Patent- und Nicht-Patentliteratur

Nachfolgend wird die Vorgehensweise zur Durchführung einer bibliometrischen Analyse beispielhaft auf der Grundlage von Patent- und Nicht-Patentliteratur vorgestellt. Bibliometrische Analysen von wissenschaftlicher Literatur sind nicht nur für den gesamten Forschungs- und Innovationsprozess von Bedeutung, sondern auch für Akteure aus Wirtschaft, Gesellschaft und Politik relevant.²³ Die bibliometrischen Indikatoren, die aus wissenschaftlicher Literatur ermittelt werden, weisen jedoch auf die Anfangsphase des Innovationsprozesses und damit vom Markt weg.

4.1 Patentinformationen als Indikatoren technisch-wissenschaftlicher Entwicklungen

Während in wissenschaftlicher Literatur der Stand von Forschung und Entwicklung dargestellt und Thesen innerhalb der Scientific Community diskutiert werden, sind Patendokumente am Stand der Technik ausgerichtet. Mit ihren praktischen Ausführungsbeispielen richten sich Patente an den sogenannten Durchschnittsfachmann. Bei Patentanmeldungen kann davon ausgegangen werden, dass ein konkretes Ergebnis aus Forschungs- und Entwicklungsarbeiten vorliegt und dass der Anmelder aufgrund des Neuheitswertes auf einen ökonomischen Nutzen hofft.²⁴ Der technische Fortschritt wird zudem in Patendokumenten in einem relativ frühen Stadium dokumentiert.

Es gibt eine Reihe von Indikatoren, die bei einer systematischen Auswertung von Patentliteratur herangezogen werden können. So kann beispielsweise die Gesamtzahl der in einem Bereich tätigen Anmelder und Erfinder als Früherkennungsindikator gewertet werden: Bei einer wachsenden Anzahl von Anmeldern und Erfindern sowie einer erhöhten Patentaktivität in einem begrenzten Zeitraum liegt die Vermutung nahe, dass innerhalb von bestimmten Technologiefeldern neue Problemlösungen erarbeitet werden konnten.

²³ In diesem Zusammenhang wird von einer zunehmenden „Verwissenschaftlichung der Gesellschaft“ sowie einer „Vergesellschaftung der Wissenschaft“ gesprochen. Zur Bedeutung des Stellenwerts bibliometrischer Wissenschaftsindikatoren vgl. [7].

²⁴ „Patente werden für Erfindungen erteilt, die neu sind, auf einer erfinderischen Tätigkeit beruhen und gewerblich anwendbar sind.“ § 1 PatG – Patentfähige Erfindung, URL: <http://bundesrecht.juris.de/patg>.

4.2 Konkurrenzanalysen zwischen Industrieunternehmen

Mit Hilfe einer Konkurrenzanalyse kann das Profil eines Unternehmens oder einer Forschungseinrichtung mit den anderen am Markt tätigen Institutionen verglichen werden. Die Beobachtung von Mitbewerbern ist unabdingbar, da in den unterschiedlichen Marktsegmenten oder Tätigkeitsbereichen verschiedene Anbieter gegeneinander konkurrieren. Eine Konkurrenzanalyse ist also ein Instrument, um sich am Markt zu behaupten oder Kollisionen zu vermeiden.

Zu Beginn einer derartigen Analyse erfolgt die Zusammenstellung von Stamm- oder Grunddaten der Konkurrenten (z. B. Umsatzgröße, Mitarbeiterzahl, Marktanteile oder Preisstrategie). Darüber hinaus müssen unternehmensrelevante Umfeld-Entwicklungen frühzeitig identifiziert werden. Mit Hilfe einer Analyse sollen Antworten auf typische Fragestellungen geliefert werden, beispielsweise auf die Fragen, welche Betätigungsfelder lukrative Entwicklungspotentiale bieten, welche Techniken in den nächsten Jahren anwendungsreif oder welches die wichtigsten Absatzmärkte für technische Lösungen sind.

4.3 Analyse und ihre Instrumente

STN[®] AnaVist[™] ist ein Analyse- und Visualisierungstool, das auf die CAS-Datenbank aufsetzt [21]. Diese Datenbank ist die umfassendste Sammlung von Veröffentlichungen aus dem Bereich der Chemie. Eine Recherche läuft mit Hilfe dieser Software im wesentlichen in folgenden Schritten ab: In einem ersten Schritt müssen die Daten gesammelt werden, d. h. über eine Suchanfrage wird ein Set von Dokumenten zusammengestellt (collect). Danach kann die Visualisierung der Daten beginnen (visualize). Dabei werden unterschiedliche Darstellungsformen systemseitig zur Verfügung gestellt (z. B. Balkendiagramme, Matrizen, Landscape). Eine Reihe von Features unterstützt das Erkunden der Daten (explore). So können beispielsweise über Filterfunktionen Daten ein- oder ausgeblendet sowie Diagramme vergrößert und damit Details herausgezogen werden; einzelne Datensätze können farblich markiert und sortiert werden. AnaVist[™] bietet vielfältige Möglichkeiten zum lokalen Speichern und Drucken der Ergebnisse (save and print).

4.3.1 Diagramme und Matrizen

Für die Rechercheergebnisse wird die Häufigkeit des Vorkommens von bestimmten Ergebnissen ermittelt. Häufigkeitsverteilungen beantworten beispielsweise die Fragen, wieviele Patente ein Erfinder angemeldet hat oder wieviele Patente in einem bestimmten Zeitraum angemeldet wurden. Nachfolgend werden zwei aufbereitete Datensätze vorgestellt, die mit Hilfe eines

Tabellenkalkulationsprogramms erstellt wurden.²⁵ In Abbildung 10 sind Forschungseinrichtungen und die Anzahl ihrer Veröffentlichungen zusammengestellt, erfasst wurden Patent- und Nicht-Patentliteratur. Bei den Einrichtungen handelt es sich nicht ausschließlich um Hochschuleinrichtungen.²⁶

Die Forschungseinrichtungen können in drei Klassen eingeteilt werden: Einrichtungen, die ein hohes (Bereich 1), die ein mittleres (Bereich 2) und die ein geringes (Bereich 3) Publikationsaufkommen besitzen. Setzt man Publikationstätigkeit mit Forschungsintensität oder eingesetzten Forschungsmitteln gleich, sind die Einrichtungen aus dem Bereich 1 *intensiv*, die Einrichtungen aus dem Bereich 2 *gelegentlich* und die aus dem Bereich 3 *selten* in dem betrachteten Forschungsbereich tätig.

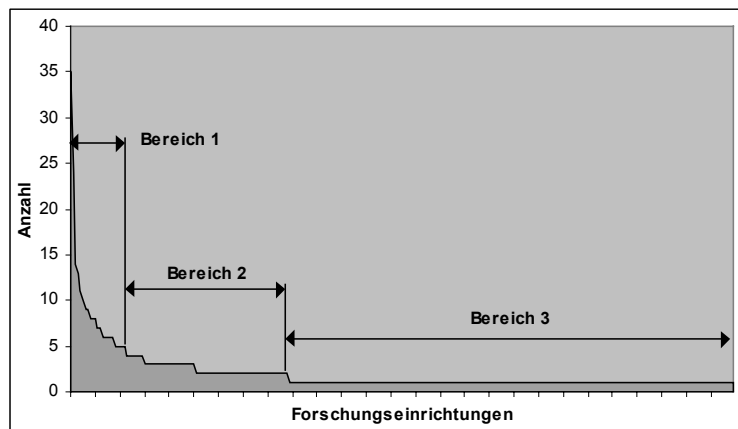


Abb. 10: Forschungseinrichtungen und ihre Publikationstätigkeit

Mit Hilfe von STN[®] AnaVist[™] können insbesondere Publikationen analysiert werden, denen sogenannte Technologie-Indikatoren zugeordnet sind. In der nachfolgenden Abbildung 11 ist die Entwicklung dieser Indikatoren für die Zeitsegmente 1975-1990, 1975-2000 und 1975-2006 aufgetragen. Ein Vergleich dieser drei Segmente zeigt, dass insgesamt ein Anstieg der Publikationstätigkeit in den Jahren 1975 bis 2006 zu verzeichnen ist. Der Indikator „Ceramics“ liegt mit großem Abstand im letzten Zeitsegment an der Spitze, es folgen leicht absteigend die anderen Indikatoren. Im Gegensatz zum

²⁵ Die Rechercheergebnisse können über AnaVist[™] im CSV-Format auf dem lokalen Rechner gespeichert werden.

²⁶ Zur Erläuterung der Graphik: Die Namen der jeweiligen Institutionen sind im Diagramm aus Platzgründen nicht aufgeführt.

Zeitsegment 1975 bis 2006 zeigt das mittlere Zeitsegment 1975 bis 2000 deutliche Sprünge zwischen den einzelnen Indikatoren. Dies könnte ein Hinweis darauf sein, dass sich in den zurückliegenden Jahren die Forschungsaktivitäten insgesamt verbreitert haben, also neue Technologiebereiche für Unternehmen an Interesse gewonnen haben.

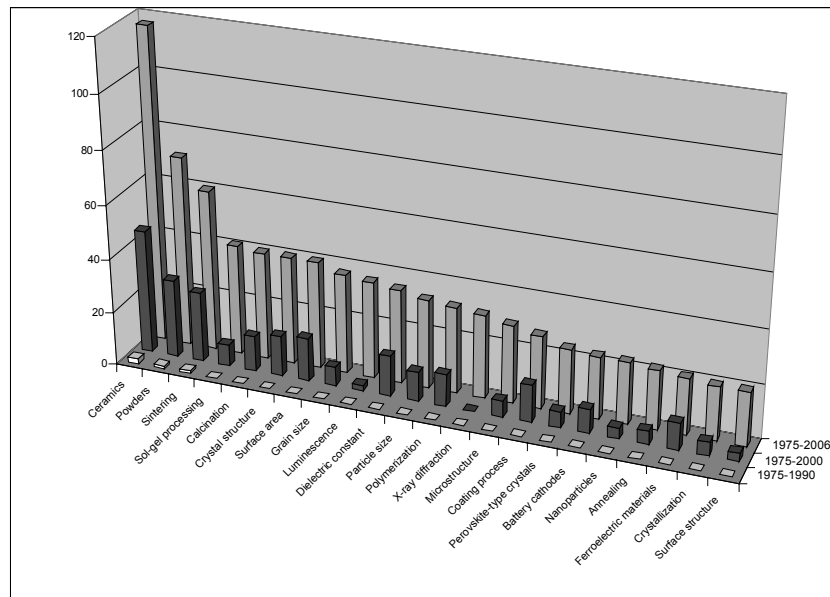


Abb. 11: Technologie-Indikatoren in unterschiedlichen Zeitsegmenten

Rechercheergebnisse können auch in Form von Matrizen dargestellt werden. Es gibt eine Reihe von unterschiedlichen Auswertungsmöglichkeiten: Die Verteilung von Patent- und Nicht-Patentliteratur, die Häufigkeit von Erfindern oder die Häufigkeit von Anmeldern wird über die Software AnaVist™ in sogenannten Charts zusammengefasst. Diese eröffnen für Nutzer vielfältige Interaktionsmöglichkeiten.

4.3.2 Research Landscape

Auf einer topographischen Karte können Recherchedaten ausgebreitet werden. Jedes Dokument ist als Punkt innerhalb dieser Landschaft wiederzufinden. Erhebungen in der Karte, also Berge, sind als zusammenhängende Cluster von Dokumenten mit ähnlichem Inhalt gekennzeichnet. Die Höhe der Gebirgszüge macht eine Aussage über die Häufigkeit der Suchergebnisse; die räumliche

Nähe der Gebirgszüge steht in Beziehung zur inhaltlichen Nähe zwischen den Dokumenten (vgl. Abb. 12). Ähnliche Inhalte werden auf der Grundlage einer Analyse von signifikanten Begriffen (Terms) der Titel- und der Abstractdaten herausgearbeitet. Beim Überstreichen der Landschaft mit dem Cursor werden weitere Begriffe in der Research Landscape mit der Häufigkeit ihres Vorkommens angezeigt.

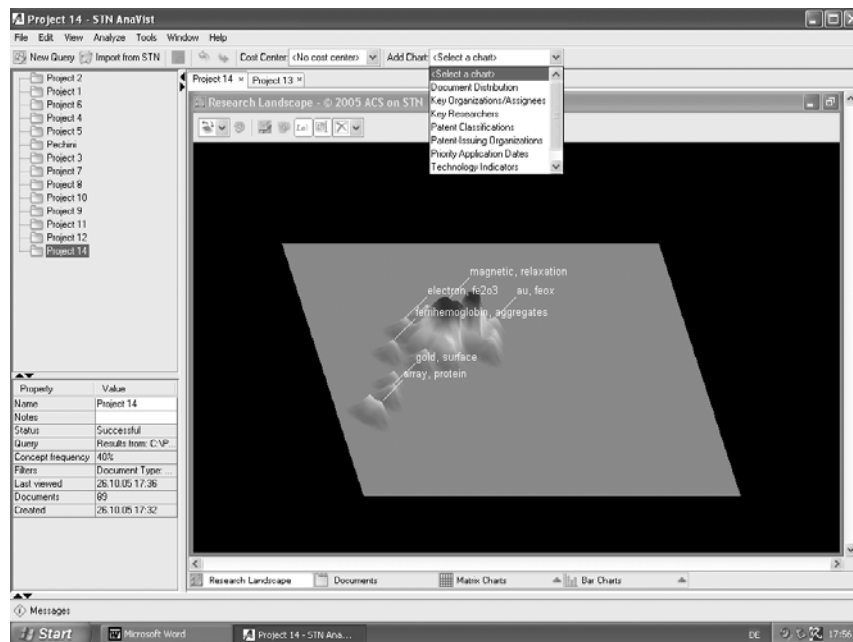


Abb. 12: Projektverwaltung und Visualisierung mit Hilfe von STN[®] AnaVist[™]

Das System bietet darüber hinaus eine Reihe von ausgefeilten Funktionalitäten und Interaktionsmöglichkeiten: Dokumente können markiert und farblich hervorgehoben, Bereiche der Research Landscape können vergrößert und unterschiedliche Perspektiven auf die Karte gewählt werden. Die Bildschirm-Oberfläche der Software AnaVist[™] ist übersichtlich gestaltet, sie ist in drei Bereiche unterteilt: Verwaltung der Rechercheergebnisse in Projektordnern (linke Bildschirmleiste), Darstellung der Rechercheergebnisse beispielsweise in Form einer Research Landscape (Hauptfenster) und Anordnung der Schaltflächen in einer Menüleiste (obere Bildschirmfläche).

5 Visualisierungspipeline und Module eines Visualisierungstools

Die auf dem Markt befindlichen Systeme bieten eine Reihe von unterschiedlichen Funktionalitäten an. Bei aller Unterschiedlichkeit in ihrem Funktionsumfang können drei wesentliche Module für die Visualisierungssysteme identifiziert werden [15]. Die Rechercheergebnisse durchlaufen diese Module, d. h. Daten werden nacheinander, wie in einer Pipeline, verarbeitet (vgl. Abb. 13). Zunächst muss die Suchanfrage durch eine Suchmaschine bearbeitet und eine strukturierte Liste von Ergebnissen ausgegeben werden. Die Ergebnisse durchlaufen danach ein Filtermodul; unterschiedliche Filterstufen können realisiert sein. Die so aufbereitete Datenmenge wird über das Modul Mapping in ein internes Darstellungsmodell umgewandelt. Der Renderer setzt dieses interne Modell in eine zweidimensionale Bildschirmdarstellung um.

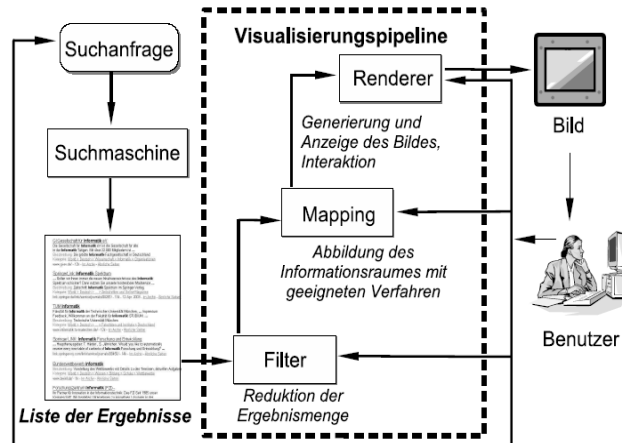


Abb. 13: Visualisierungspipeline

Die am Bildschirm generierte Darstellung ist Grundlage für eine Änderung der Suchanfrage bzw. für eine Änderung der graphischen Darstellung. Der Eingriff in die Visualisierungspipeline kann – je nach Leistungsumfang der Software – an den genannten Modulen erfolgen. Es können beispielsweise Einstellungen an den Filtern vorgenommen, unterschiedliche Darstellungsformen der Rechercheergebnisse gewählt oder Interaktionen mit der Darstellung durchgeführt werden. Visualisierung ist damit nicht nur eine Erweiterung der allgemein üblichen listenweise Ausgabe von Rechercheergebnissen, sondern liefert darüber hinaus die Möglichkeit, an unterschiedlichen Stellen der Informationspipeline mit Daten zu interagieren.

6 Zusammenfassung und Ausblick

Über das Internet wird eine Vielzahl an unterschiedlichen Quellen bereitgestellt.²⁷ Mit Hilfe von Visualisierungsinstrumenten können Strukturen aus umfangreichen Datenmengen herausgearbeitet und Beziehungen zwischen Informationseinheiten aufgezeigt werden. Die in Kapitel 2 beispielhaft vorgestellten softwaretechnischen Instrumente zeigen, dass mit ihrer Hilfe die Datenvielfalt und -komplexität reduziert werden kann. Die Analysebeispiele aus den Kapiteln 3.3 und 4.3 verdeutlichen die Arbeitsschritte, die die Daten während ihrer Aufbereitung durchlaufen. Vor dem Hintergrund des Modells einer Visualisierungspipeline können die Eingriffsmöglichkeiten auf die einzelnen Visualisierungsmodule dargestellt werden. Statt lediglich große Mengen an Informationen in einer digitalen Bibliothek bereitzustellen, muss eine Strukturierung für den Nutzer erfolgen.²⁸ Die damit verbundenen Abläufe müssen ihrerseits dem Nutzer vermittelt werden. Informationsvisualisierung kann somit nicht nur als Instrument, sondern auch als Methode verstanden werden.

Die gegenwärtigen Forschungs- bzw. Untersuchungsrichtungen im Bereich der Informationsvisualisierung zeigen, dass grundlegende Kriterien fehlen, die den Wert einer Informationsvisualisierung erfassen und damit vergleichende Beurteilungen zwischen unterschiedlichen Systemen ermöglichen. Das Urteil über eine gute oder schlechte, eine geeignete oder ungeeignete Visualisierung wird vom jeweiligen Nutzer gefällt und hängt u. a. von seiner bildlichen oder räumlichen Vorstellungskraft ab.

Die gegenwärtigen Entwicklungen aus dem Umfeld der digitalen Bibliotheken belegen, dass konventionelle und innovative Visualisierungstechniken in Literatur- und Fachdatenbanken eingesetzt werden. Des weiteren zeigen die Entwicklungen im WWW sehr deutlich, dass Daten in immer stärkerem Maße in Form von Übersichtsgraphen und Strukturnetzen angeboten werden.²⁹

²⁷ Eine Übersicht über Online-Kataloge, digitale Bibliotheken und Fachportale liefert [19].

²⁸ In diesem Zusammenhang muss auf den Vorzug eines guten Bibliothekswesens hingewiesen werden: „Die Bibliothek ist auch heute noch die wichtigste Waffenkammer gegen den Unsinn dieser Welt, um ihn aber zu besiegen muss man ihn kennen, richtig einordnen und durchschauen. Bibliographien und Kataloge als moderne Suchmaschinen sind dabei der Schlüssel zur Macht durch publiziertes Wissen.“ [25, S. 1456]

²⁹ Der Fisheye-View wird bei der Darstellung von hierarchischen Strukturen bevorzugt, vgl. [8], S. 1. Zum Graphenlayout und zur angewandten Netzwerkvisualisierung vgl. [6].

Literatur und Internetquellen

- 1 *An atlas of cyberspaces – Maps of internet service provider and internet backbone networks.* URL: http://www.cybergeography.org/atlas/isp_maps.html.
- 2 BALL, R. & TUNGER, D. (2005). *Bibliometrische Analysen – Daten, Fakten und Methoden.* Jülich: Forschungszentrum Jülich.
- 3 BATAGELJ, V. (2003). Efficient algorithms for citation network analysis. *Computer science, abstract. cs.* DL/0309023. URL: http://arxiv.org/PS_cache/cs/pdf/0309/0309023.pdf.
- 4 *BibTechMon.* URL: <http://www2.arcs.ac.at/S/ST/BTM/bibtech.engl.pdf>.
- 5 BÖRNER, K. [u. a.] (Hrsg.). (2002). *Visual interfaces to digital libraries: Motivation, utilization, and socio-technical challenges.* Berlin: Springer.
- 6 BRANDES, U. & WAGNER, D. (2004). Netzwerkvisualisierung. *it-Information Technology* 46, 3, S. 129-134.
- 7 CENTER FOR SCIENCE & TECHNOLOGY STUDIES BERN. (1999). *Wissenschaft und Innovationsprozess: Entwicklungen, Konzepte und Indikatoren.* URL: http://www.cest.ch/Publikationen/1999/am/fop_99_55/kapitel_1.pdf
- 8 CHEN, CH. (2004). *Information visualization – Beyond the horizon.* 2nd ed. London [u. a.]: Springer.
- 9 GARFIELD, E. (1955). Citation indexes for science. *Science* 122, 3159, S. 108-111.
- 10 *GRIDL – GRaphical Interface for Digital Libraries.* URL: <http://www.cs.umd.edu/hcil/west-legal/gridl>.
- 11 *HISTCITE™ – Bibliographic Analysis and Visualization Software.* URL: <http://www.histcite.com/>.
- 12 HUMMON, N. P. & DOREIAN, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social networks* 11, S. 39-63.
- 13 KELLER, T. & TERGAN, S.-O. (2005). Visualizing knowledge and information: An introduction. In Tergan, S.-O. [u. a.] (Hrsg.). *Knowledge and information visualization – Searching for synergies* (S. 1-23). Berlin: Springer.
- 14 *Mooter-Web-Search.* URL: <http://www.mooter.com>.
- 15 NAJARRO, M. A. BARBERENA (2003). *Visualisierung von Informationsräumen.* URL: http://www.bibliothek.tu-ilmenau.de/elektr_medien/dissertationen/2003/Barberena%20Najarro_Martha/v.pdf.

- 16 NOURMOHAMMADI, H. A. & UMSTÄTTER, W. (2004). Die Verteilung der Autorenhäufigkeit in wissenschaftlichen Zeitschriften bei verschiedenen Themen und Ländern. *Information – Wissenschaft & Praxis* 55, 5, S. 267-274.
- 17 PRICE, D. J. DE SOLLA. (1965). Networks of scientific papers. *Science* 149, 3683, S. 510-515.
- 18 SCHWARTZ, D. (1999). Mehr Information durch Visualisierung von Daten? Konventionelle und innovative Visualisierungstechniken. *BIT-Online* 2, 4, S. 455-462. URL: <http://www.b-i-t-online.de/archiv/1999-04/nach1.htm>.
- 19 SCHWARTZ, D. (2005). *Digitale Bibliotheken und Portale. Informations- und Dienstleistungsangebote in der Wissensgesellschaft*. Hamburg: Dashöfer.
- 20 *SciFinder Scholar*. URL: <http://www.cas.org/SCIFINDER/SCHOLAR/index.html>.
- 21 *STN AnaVist - Analyze and visualize online search results for better decision making*. URL: <http://www.cas.org/stnavist/>, http://www.stn-international.de/stninterfaces/stnavist/stn_anavist.html.
- 22 UMSTÄTTER, W. (1995). Die Rolle der Digitalen Bibliothek in der modernen Wissenschaft. *Humboldt-Spektrum* 2, 4, S. 36-41. URL: <http://www.ib.hu-berlin.de/~wumsta/pub93.html>.
- 23 UMSTÄTTER, W. (1998). Über die Messung von Wissen. *Nachrichten für Dokumentation* 49, 4, S. 221-224. URL: <http://www.ib.hu-berlin.de/~wumsta/infopub/pub1996f/dhb3.html>
- 24 UMSTÄTTER, W. (1999). *Zum Thema: Von Little Science zu Big Science*. [Rez. zu Price, D. J. de Solla. (1974). *Little Science, Big Science*. Frankfurt am Main: Suhrkamp.] URL: <http://www.ib.hu-berlin.de/~wumsta/price14.html>.
- 25 UMSTÄTTER, W. (2005). Bibliographien, Kataloge, Suchmaschinen. Das Ende der Dokumentation als modernes Bibliothekswesen. *Bibliotheksdienst* 39, 11, S. 1442-1456. URL: http://www.zlb.de/aktivitaeten/bd_neu/heftinhalte2005/Bibliotheken011105.pdf.

Die zitierten Internetquellen wurden zuletzt am 27.07.2006 aufgerufen.