

Bibliometrie als DataMining-Werkzeug in der Naturwissenschaft

Dirk Tunger, Jülich

Abstract

Dieser Aufsatz veranschaulicht, wie mit Hilfe bibliometrischer Methoden grosse Datenmengen ausgewertet werden können. Es soll gezeigt werden, dass ein Informationsmehrwert nicht nur auf der inhaltlichen Ebene zu finden ist, sondern auch auf einer Metaebene interessante Informationen verborgen sind, die mit metrischen statistischen Methoden ermittelt werden können.

Es wird eine Einführung in das Thema Bibliometrie gegeben und an einem Praxisbeispiel gezeigt, wie diese theoretischen Annahmen in die Praxis umgesetzt werden können.

1. Einführung

Problemstellung

Es ist keine Neuigkeit mehr, dass die produzierten Datenmengen immer größer werden. Das Problem: Die Inhalte selbst sind zu einem sehr großen Teil nicht mehr zu bewältigen. Dies bedeutet in der Praxis: von der inhaltlichen Seite ist das Problem nicht zu lösen.

Probleme bei der Lösung

DataMining ist oft als Rettung angepriesen worden, im Arbeitsalltag von Informationsspezialisten in Bibliotheken ist davon aber bisher wenig angekommen. Sind die entwickelten Tools auf der einen Seite zu speziell, verlangen sie zu viel theoretisches Wissen? Sind die DataMining-Theorien auf der anderen Seite zu allgemein? Fehlen nur best practise-Beispiele? Einige Fragen scheinen noch ungeklärt.

Lösungsweg

Der Einsatz von Bibliometrie (<http://www.bibliometrie.de>) als Werkzeug für die Datenauswertung ist nicht unmittelbar neu, die damit verbundenen Möglichkeiten scheinen aber bisher nur sehr selten genutzt zu werden. Bibliometrie ist ein Wissenschaftszweig, in dessen Mittelpunkt die statistische Auswertung von wissenschaftlichen Veröffentlichungen steht (Forschungszentrum Jülich, Zentralbibliothek, 2003). Der Begriff „wissenschaftliche Veröffentlichungen“ ist dabei sehr weit gefasst. Er bezieht sich nicht nur auf Veröffentlichungen in Zeitschriften, sondern kann auch Bücher, Webseiten oder Patente einschliessen.

Ziel

Ziel ist es, Möglichkeiten aufzuzeigen, wie mit Hilfe der Bibliometrie Informationen gewonnen werden können, die mit konventionellen Methoden des Information-Retrieval nicht generiert werden können.

2. State of the Art in der Bibliometrie

Einsatzgebiete von Bibliometrie

Bibliometrie lässt sich hervorragend einsetzen, um einzelne Wissenschaftsgebiete oder wissenschaftliche Einrichtungen gezielt zu untersuchen (Ball, R; Tunger, D., 2005). Ziel einer solchen Untersuchung kann sein, die zeitliche Entwicklung eines Themas zu verfolgen:

- Wie viele Artikel wurden zu einem bestimmten Thema veröffentlicht?
- Wie hat sich dieses Publikationsverhalten im Laufe der vergangenen Jahre geändert?
- Wie ist die Resonanz auf ein Thema? Welche Änderungen hat es hier gegeben?

Ebenso lässt sich Bibliometrie für die Wissenschaftsevaluation nutzen (van Raan, A., 2004). In diesem Fall erhält man Antworten auf Fragen wie:

- Welches sind die führenden Einrichtung zu einem Thema?
- Wie wurden die Veröffentlichungen einer bestimmten Einrichtung im Vergleich mit thematisch ähnlich ausgerichteten Einrichtungen wahrgenommen?
- Welches sind die wahrnehmungsstärksten Zeitschriften einer Disziplin.

Weitere Informationen zur Bibliometrie im Fachportal www.bibliometrie.de.

Ein Bezugsrahmen muss gewählt werden

Veröffentlichungszahlen oder Zitationszahlen allein sagen recht wenig aus, wenn Sie nicht in einen vernünftigen Bezugsrahmen gesetzt werden. Aussagen können zum Beispiel getroffen werden, wenn der Bezugsrahmen zu thematisch ähnlichen Einrichtungen, zu Ländern oder zu einer ausgewählten Fachöffentlichkeit gewählt wird.

Wissenschaftskommunikation

Ein Wissenschaftler veröffentlicht im Wesentlichen aus zwei Gründen:

- Zur Problemlösung in seiner Disziplin
- Zur Erhöhung der eigenen Reputation. Dies bedeutet, der Wissenschaftler möchte durch seine Ergebnisse auf sich aufmerksam machen und seine neuen Methoden vorstellen.

In seinen Veröffentlichungen zitiert ein Wissenschaftler demzufolge, um seine eigenen Ergebnisse mit den Ergebnissen anderer Wissenschaftler zu untermauern.

Er zitiert aber auch, um auf vorausgegangene Veröffentlichungen (Ergebnisse) hinzuweisen.

Datenbasis für bibliometrische Analysen

Wissenschaftliche Veröffentlichungen existieren in verschiedenen Formen: Unter anderem in Büchern, Konferenzbänden und Aufsätzen in wissenschaftlichen Zeitschriften. Bei der Messung der Zitationshäufigkeit werden allerdings oft nur wissenschaftliche Zeitschriften beachtet. Dies liegt an der Zusammensetzung der Datengrundlage: Eine Datenbank, die unter Wissenschaftlern als Science Citation Index (SCI) bekannt ist, wertet regelmässig etwa 5900 naturwissenschaftliche Zeitschriften aus. Dies ist der einzige multidisziplinäre Zitationsindex, der zusätzlich zu bibliographischen Angaben auch die Zitationen der Veröffentlichungen auswertet. Aus den sozialwissenschaftlichen Disziplinen kommen noch einmal etwa 1200 Zeitschriftentitel dazu. Das klingt insgesamt viel, ist es aber nicht: weltweit existieren ca. 120.000 wissenschaftliche Zeitschriften aller Disziplinen.

Ausgewertet und für bibliometrische Analysen zu Grunde gelegt werden also gerade einmal 5% der wissenschaftlichen Veröffentlichungen in Zeitschriften. Von den unzähligen Büchern und Konferenzbeiträgen werden nur die allerwenigsten erfasst. Im Umkehrschluss bedeutet dies: Auch Zitate werden damit nur aus diesen etwa 5% der wissenschaftlichen Veröffentlichungen komplett erfasst.

Standardisierung der Naturwissenschaft

Für die Naturwissenschaften bestehen bei der Datenauswahl keine Probleme, da diese sehr international ausgerichtet sind: Naturwissenschaftliche Themen sind weltweit von Interesse, die Fragestellungen ähneln sich. Kommunikationssprache ist Englisch und der größte Anteil naturwissenschaftlicher Arbeiten erscheint in Form von Aufsätzen in Zeitschriften. Bücher spielen in den Naturwissenschaften nur eine untergeordnete Rolle.

Man kann sagen, in den Naturwissenschaften herrschen weltweit sehr ähnliche Standards. Dies ist ein grosser Vorteil und ermöglicht auch erst internationale Vergleiche.

In den Geisteswissenschaften sieht es hingegen anders aus: Themen sind teilweise nur von eingeschränkter regionaler Bedeutung und oftmals in Nationalsprachen abgefasst. Daraus entsteht ein Problem: Für internationale Journals ist das Interesse an derartigen Aufsätzen gering, vor allem, wenn der Bezug zu den USA fehlt. Damit ist es schwierig, in den internationalen sozialwissenschaftlichen Journals Beiträge unter deutscher Beteiligung zu platzieren.

Für die Sozialwissenschaften sind damit die Möglichkeiten internationaler Vergleiche nur schwer anwendbar. Hinzu kommt, dass Bücher eine wesentlich größere Bedeutung einnehmen als in den Naturwissenschaften.

Bildung von Indikatoren

Mit Hilfe von Indikatoren kann eine große Anzahl an Veröffentlichungen beurteilt werden. Hierbei findet keine qualitative Beurteilung statt, sondern eine quantitative. Ein Indikator kann beispielsweise die Anzahl der Zitate pro Artikel (Zitationsrate) benennen. Bei der Bildung dieses Indikators ist dies weniger für einen Artikel interessant, als vielmehr für ein Set an Artikeln. Dieses Set kann dann zu Vergleichs-Sets in Bezug gesetzt werden. Auf diese Weise entsteht ein Ranking, das aus verdichteten Daten besteht und einen Überblick in der Bewertung der untersuchten Artikel liefert.

Ebenso kann man mittels Bibliometrie Vernetzungen und Interdisziplinarität in der Wissenschaft aufzeigen.

Ziel bei der Bildung von Indikatoren ist es, eine vergleichbare Umgebung zu erzeugen, die Grössenunterschiede wissenschaftlicher Einrichtungen relativiert.

3. Kombination von DataMining und Bibliometrie

„DataMining ist die Gewinnung impliziter, bislang unbekannter und potenziell nützlicher Informationen aus Daten“ (Witten, I.; Eibe, F., 2001).

Für die professionelle Anwendung von DataMining existieren etliche Programme, die Unterstützung bieten sollen. Ein sehr bekanntes Tool ist „WEKA“ von der neuseeländischen Universität Waikato. Die Erzielung von brauchbaren Ergebnissen mit diesen Programmen hängt aber von der Struktur der Daten und den Kenntnissen der Auswertungsalgorithmen dieser Programme ab.

Aber auch Abseits von speziellen DataMining-Tools lässt sich DataMining betreiben: Bibliometrie ist zwar gleich ein kompletter Wissenschaftszweig, er hat aber genau das oben beschriebene Ziel vor Augen: die Gewinnung von Informationen aus Daten. Die Schwierigkeiten sind die gleichen, die auch beim tool-unterstützten DataMining auftreten: Die Aufbereitung von Daten vor der weiteren Analyse ist zeitaufwendig.

Der Einsatz bibliometrischer Methoden lohnt sich vor dem Hintergrund, eine grosse Anzahl an wissenschaftlichen Aufsätzen auf einmal auszuwerten und daraus die gewünschten zusätzlichen Informationen zu ziehen.

Daten zu Information veredeln – so kann man kurz das Hauptziel von DataMining fassen (Grötter, R., 2002).

4. Bibliometrie als Teil eines Trenderkennungssystems

Bibliometrie ist mehr als nur ein Werkzeug, mit dem Wissenschaftsevaluation betrieben werden kann. Bibliometrie kann als Controlling-Instrument in der Wissenschaft auch zur Trenderkennung genutzt werden.

Es muss an dieser Stelle bemerkt werden, dass Bibliometrie natürlich nur einen Teilbereich in einem Trenderkennungssystem bildet. Neben wissenschaftlichen Veröffentlichungen sind Patente und Konferenzveröffentlichungen für die technologische Entwicklung ebenfalls relevant.

Neben technologischer Entwicklung sind in einem Trenderkennungssystem auch noch weitere Ebenen von Bedeutung (Gomez, P., 1983).

Die

- soziale Ebene
- politische Ebene
- ökonomische Ebene und
- technologische Ebene

müssen in einem Trenderkennungssystem eine Einheit bilden (Rieser, I., 1980).

Blick zurück nach vorn

Wie auch an anderer Stelle, steigt in der Welt der Wissenschaft die Zahl an verfügbaren Inhalten (wissenschaftliche Veröffentlichungen). Die Zeit, einzelne Ergebnisse wahrzunehmen, wird immer geringer. Dadurch wird nur ein Bruchteil der erzielten wissenschaftlichen Ergebnisse intensiv gelesen und weiterverarbeitet.

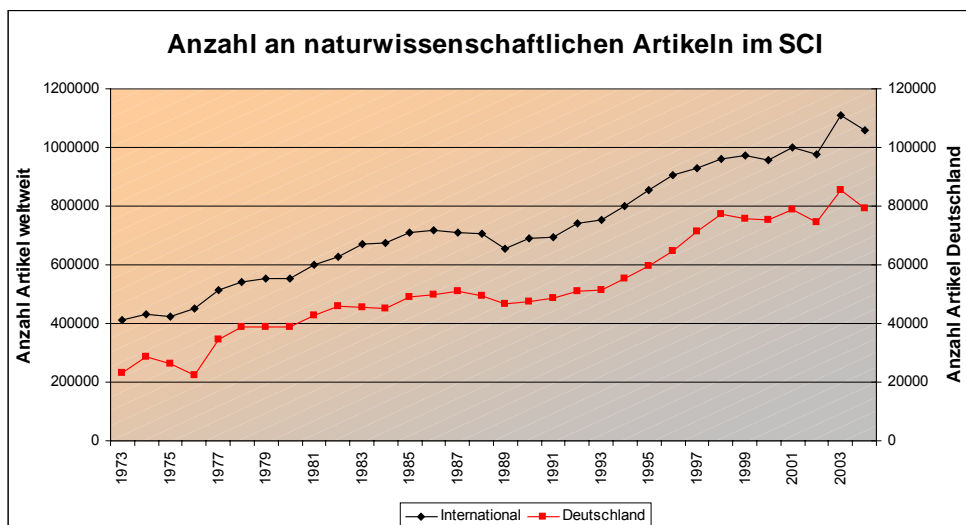


Abbildung 1: Darstellung des zeitlichen Verlaufs der Produktion wissenschaftlicher Artikel (nur im Science Citation Index verzeichnete)

Das Problem liegt darin, einen Überblick über den bisherigen Stand der Forschung zu erhalten. Nur mit einem Überblick ist es aber möglich, ergebnisorientiert und effizient zu forschen. Eine praktisch anwendbare Möglichkeit in der Wissenschaft,

einen Trendscout aufzustellen, der als Trenderkennungssystem fungiert, ist die Durchführung von bibliometrischen Analysen

Mit bibliometrischen Analysen lässt sich beispielsweise die technologische Entwicklung von wissenschaftlichen Disziplinen messen. Die Vorgehensweise ist vom Arbeitsaufwand her vertretbar und vom Ergebnis her sehr aussagekräftig: In der Datenbank „Science Citation Index“ wird zu einem Grossteil wissenschaftlicher Veröffentlichungen die Anzahl der Zitationen verzeichnet.

Aus diesem Grund ist es möglich, zu einer thematisch ausgerichteten Recherche Antworten auf unter anderem drei Fragen zu erhalten:

Vergangenheits-Aspekt: Wie hat sich die Anzahl wissenschaftlicher Veröffentlichungen über einen bestimmten Zeitraum entwickelt?

Gegenwarts-Aspekt: Wie wurden diese Artikel zitiert?

Zukunfts-Aspekt: Gab es bei den Publikationszahlen/ Zitierungen grosse Zuwächse oder Einbrüche?

Der **Vergangenheits-Aspekt** wird gebildet von der *Anzahl der Artikel* in der Datenbank zu einem Thema. Die Betrachtung der Anzahl von Veröffentlichungen schaut in die Vergangenheit, da diese Zahl für den betrachteten Zeitraum nicht mehr zu verändern ist.

Der **Gegenwarts-Aspekt** wird gebildet aus der Anzahl an Zitationen auf die existierenden Artikel. Diese Zahl kann sich täglich ändern, wenn einer der Artikel in anderen Veröffentlichungen zitiert wird.

Der **Zukunftsaspekt** wird gebildet aus den Zuwächsen oder Einbrüchen der Publikationen und Zitationen über einen längeren Zeitraum. Die Veränderung gegenüber einer Vorperiode lässt erkennen, ob das wissenschaftliche Interesse zu- oder abnimmt.

Der Zukunfts-Aspekt ist demnach der wichtigste von den drei Aspekten. Es sind nicht die absoluten Zahlen an Artikeln oder Zitationen, die die Zukunftsaussage tragen, sondern die Veränderung der Zahl an Zitationen im Verhältnis zum Jahr davor.

Man kann einwerfen, dass auch der Zukunftsaspekt auf Zahlen aus der Vergangenheit basiert. Dies ist auch richtig, doch ergibt sich dieses Problem an jeder Stelle, wo von Zukunft die Rede ist. Es ist schlicht unmöglich, Zahlen aus der Zukunft zu erhalten. Aus diesem Grund müssen die ermittelbaren Zahlen in ein Verhältnis gebracht werden, dass sich Zahlen mit Aussagekraft in die Zukunft ergeben. Eine blosse Interpolation wäre wenig sinnvoll: Das Problem ist die Vorhersage von Wendepunkten. Würde man also Interpolation für den Aufbau eines strategischen Radars einsetzen, würde man ein falsches Gefühl der Sicherheit erzeugen.

Die Methode ist für die drei weiteren Ebenen (soziale, politische und ökonomische Ebene) zu übertragen.

Der Vorteil in der vorgestellten Methode liegt darin, immer ganz konkret Entwicklungsträger zu identifizieren. Der zweite Schritt ist die Messung von Resonanz auf diese Entwicklungsträger und die entsprechende Veränderung.

Mit einem einzelnen Indikator kann mit Sicherheit keine Vorhersage von zukünftiger Anwendbarkeit gemacht werden. Der Verbund und die geschickte Kombination mehrerer Entwicklungsträger lässt aber durchaus vernetzte Aussagen zu.

Praxisbeispiel

Die Theorie soll an einem konkreten Beispiel aus der Praxis durchgespielt werden.

Nanotechnologien / Nanomaterialien

Der Begriff geht auf Norio Taniguchi (1974) zurück und beschreibt die Entwicklung von Materialien, die in mindestens zwei Dimensionen kleiner als 100 Nanometer sind. Die Theorie besagt, dass diese Materialien ihre Eigenschaften und ihre Struktur ändern.

Technologische Ebene Die Betrachtung der wissenschaftlichen Artikel zu diesem Thema ist eindeutig: Wurden im Zeitraum 1995 – 1999 weltweit 3190 Artikel zu diesem Thema veröffentlicht, waren es im Zeitraum 2000 – 2004 bereits 9823. Die Zahl der Länder, die sich für dieses Thema interessieren, stieg von 59 auf 80.

China konnte dabei den Anteil seiner Forschung von 9 % auf 17 % fast verdoppeln, während in den USA der Anteil von 34 % auf 28 % zurückging. Deutschland hat ebenfalls einen Rückgang zu verzeichnen, von 13 % auf 10 %. Die Prozentzahlen beziehen sich auf den Anteil der Artikel in der jeweiligen Zeitperiode am weltweiten Output zu diesem Thema. Es hat somit keinen Rückgang der Artikelproduktion oder der Forschung in diesen beiden Ländern gegeben, sondern eine überproportionale Steigerung des Outputs anderer Länder (Weidenfeld, W; Turek, J., 2002). Für einzelne Themenaspekte könnten weitere Untersuchungen angestellt werden, beispielsweise, wie stark das wissenschaftliche Interesse gestiegen ist (gemessen in Form von Zitaten).

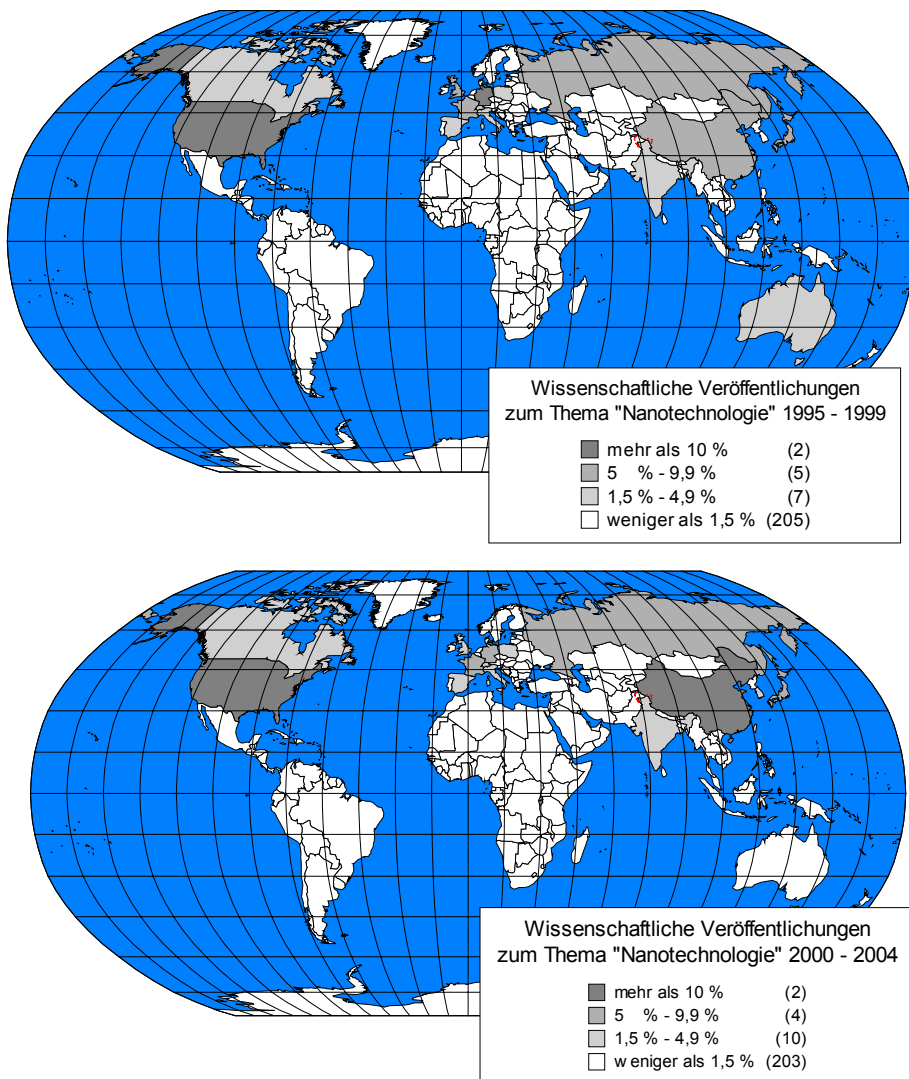


Abbildung 2: Wissenschaftliche Veröffentlichungen zum Thema „Nanotechnologie“. Darstellung des prozentualen Anteils einzelner Länder in zwei Zeitperioden

Ausblick:

Es wurde bereits angedeutet, dass nicht nur die technologische Entwicklung von Bedeutung ist für die Trenderkennung, sondern dass auch weitere Faktoren mit einfließen müssen. Für die soziale Ebene soll das Beispiel weiter ausgebaut werden.

Soziale Ebene Wie stark wird die Technologie in der Öffentlichkeit wahrgenommen? Wie oft und in welchen Massenmedien wird über Nanotechnologie berichtet?

Der Grafik liegt eine Auswertung der Datenbank GBI (Themenrubrik „Tages- und Wochenpresse“) zu Grunde.

Das Diagramm zeigt, dass eine sehr lange Zeit in der Öffentlichkeit dieses Thema nicht diskutiert wurde. Seit 2000 hat sich dies geändert: Nahezu schlagartig wurde immer öfter und weiter gestreut berichtet. Ein eindeutiges Zeichen für Interesse an diesem Themengebiet.

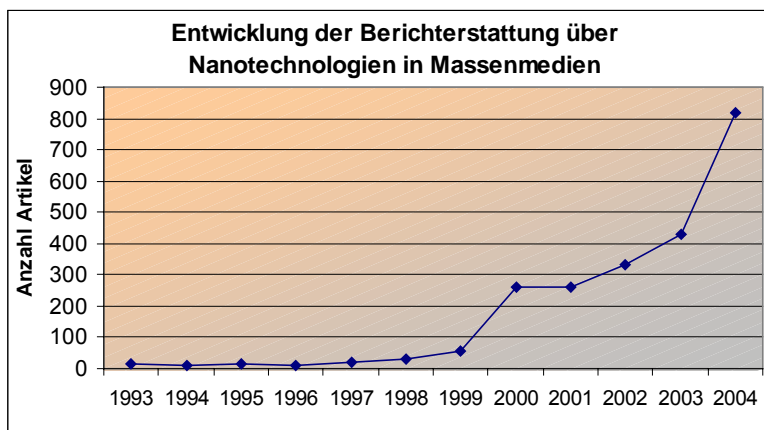


Abbildung 3: Entwicklung der Berichterstattung über das Thema „Nanotechnologie“ in täglich und wöchentlich erscheinenden Massenmedien

Mit metrischen statistischen Methoden können weitere Auswertungen gemacht werden, die Auskunft über folgende Fragen liefern:

Sind Widerstände erkennbar (Bürgerinitiativen oder Vereinsgründungen)? Ist Unterstützung zu erwarten? Derartige Einschätzungen sind durch eine standardisierte Klassifikation und Bewertung von Medienberichten zu erhalten.

Die Betrachtung eines Themas aus verschiedenen Perspektiven hilft, frühzeitig Chancen und Risiken zu erkennen. Dabei reicht eine einmalige Anstrengung nicht aus, in regelmässigen Zeitabständen müssen Veränderungen immer wieder überprüft werden. Die hier vorgestellten Ergebnisse und Methoden sind auch nur als beispielhaft zu verstehen und müssten für ein spezielles Themengebiet weiter konkretisiert werden.

Die Einbeziehung von Experten vereinfacht die Interpretation der Ergebnisse.

Integration von Trenderkennung in die Wissenschaft

In Forschungseinrichtungen sind ein besseres Controlling und eine bessere Steuerung nur über eine stärkere Nutzung bisher ungenutzter Datenbestände zu

erzielen. Die Datenbestände müssen durch Analyse- und Bewertungsverfahren zu Informationen und Wissen umgeformt werden.

Für den Bereich Wissenschaft kann hierzu die bibliometrische Analyse genutzt werden, für den Bereich Gesellschaft/ Soziales kann dies eine Medienresonanzanalyse sein.

Diese Medienresonanzanalyse kann auf Basis ähnlicher statistischer Verfahren die Tages- und Wochenpresse auswerten. Es können Aussagen über den Umfang und die zeitliche Abfolge zu einem Thema gemacht werden. Ebenso kann die Art des Mediums (Reichweite, Meinungsführerschaft etc.) in derartige Analysen miteinfließen. Eine inhaltliche Bewertung (bezüglich negativer oder positiver Berichterstattung) hilft, die Einschätzung zu verbessern.

Der Bereich Politik spielt mit in den Bereich Gesellschaft hinein und wird teilweise auch über die Massenmedien ausgedrückt. Gesetzgebungsverfahren werden vorher oft in der Presse diskutiert und deren Richtung mit beeinflusst.

Daten zum Bereich Ökonomie werden täglich in riesigen Mengen produziert. Es existieren Daten auf Ebene der Finanzmärkte, aber auch auf ausserbörslicher Ebene. Diese Daten dürfen nicht ohne Beachtung bleiben: was sagen unterschiedliche Experten über die Entwicklung einer Branche aus? Aus welcher Perspektive werden diese Voraussagen getroffen? Welche Daten liegen diesen Bewertungen zu Grunde?

5. Fazit

Es sind immer drei Dinge, die gesucht werden:

1. Träger von Entwicklungen (z.B. wissenschaftliche Veröffentlichungen)
2. Resonanz auf diese Entwicklungsträger (z.B. Zitationen)
3. Veränderung der Resonanz

Kann man entsprechende Datenquellen benennen, die diese drei Faktoren ermittelbar machen, so können entsprechende Indikatoren gebildet werden. Nicht nur das einmalige Erheben der Indikatoren ist der Informationsmehrwert, sondern eine kontinuierliche Beobachtung und eine tiefgehende Interpretation. In letzterem Punkt liegt eine nicht zu unterschätzende Schwierigkeit der Analysen.

In den beschriebenen Methoden wird DataMining nicht in der Form angewandt, dass mit Hilfe von DataMining-Tools Zusammenhänge zwischen den Datensätzen gesucht werden. Vielmehr wird in diesem Fall versucht, eine höhere Aggregationsebene zu erreichen und Daten zu Informationen umzuwandeln, durch das Clustern einer bestimmten Anzahl an Artikeln und das zusammenhängende Auswerten bestehender Datenfelder.

Literatur

- Ball, Rafael; Tunger, Dirk: Bibliometrische Analysen – Daten, Fakten und Methoden. Grundwissen Bibliometrie für Wissenschaftler, Wissenschaftsmanager, Forschungseinrichtungen und Hochschulen; Forschungszentrum Jülich, Zentralbibliothek, Reihe Bibliothek Bd. 12, 2005, ISBN: 3-89336-383-1
- Forschungszentrum Jülich, Zentralbibliothek: Bibliometric Analysis in Science and Research. Applications, Benefits and Limitations; 2nd Conference of the Central Library; Forschungszentrum Jülich, Zentralbibliothek, Reihe Bibliothek Bd. 11, 2003, ISBN: 3-89336-334-3
- Gomez, Peter: Frühwarnung in der Unternehmung. Haupt, Bern, 1983
- Grötke, Ralf: Goldgräber in der Datenmine in: Die Zeit 16/2002;
http://zeus.zeit.de/text/archiv/2002/16/200216_t-data-mining.xml
- Rieser, Ignaz: Frühwarnsysteme für die Unternehmungspraxis. Florentz, München, Wirtschaftswissenschaftliche Forschung und Entwicklung, 1980
- van Raan, Anthony: Measuring Science *in*: Moed, H.F.; Glänzel, W; Schmoch, U: Handbook of Quantitative Science and Technology Research. The Use of Publication and Patent Statistics in Studies of S&T Systems; Kluwer Academic Publishers, Dordrecht, 2004, ISBN: 1-4020-2702-8
- Weidenfeld, W; Turek, J: Wie Zukunft entsteht. Größere Risiken – weniger Sicherheit – neue Chancen; Gerling Akademie Verlag, München, 2002, ISBN: 3-932425-46-4
- Witten, Ian; Frank, Eibe: Data Mining. Praktische Werkzeuge und Techniken für das maschinelle Lernen; Carl Hanser Verlag, München, 2001; ISBN: 3-446-21533-6