

Data Mining in Literaturdatenbanken

Methoden zur Generierung von Hypothesen für die biomedizinische Forschung

Johannes Stegmann, Berlin

Der Vortrag gibt eine kurze Einführung in die Themen „Data Mining“ und „Wissensentdeckung“. Eingehender wird die Methode von Don Swanson erläutert, Verbindungen zwischen - in Bezug aufeinander - isolierten Literatursets herzustellen und so zur Aufstellung „neuer“ (unveröffentlichter) Hypothesen zu gelangen. Weiterhin werden einige Experimente für ein strukturiertes Auffinden derartiger „Literatur-Partner“ vorgestellt.

A short introduction into the subjects Data Mining and Knowledge Discovery will be given, with emphasis on Don Swanson's special method of linking disparate literatures and hypothesis generation. A preliminary result with respect to the structure of an intermediate literature will be presented, showing that it might be possible to bring related but unconnected themes in close neighbourhood using Co-Word analysis.

Einleitung

Der Begriff *Data Mining* wird häufig in Verbindung mit dem Terminus *Knowledge Discovery* („Wissensentdeckung“) bzw. *Knowledge Discovery in Databases* (KDD) verwendet. Knowledge Discovery bedeutet zunächst nichts anderes, als sinnvolle Informationen aus Daten zu extrahieren. Data Mining meint hier die Anwendung spezifischer Algorithmen zur Erkennung von Mustern in den Daten [1].

Im Zusammenhang mit Literaturdatenbanken (also Datenbanken, deren „Daten“ im wesentlichen aus Text bestehen), spricht man auch vom *Text Mining* als der Extraktion nützlicher Information aus der Literatur [2]; dieser Begriff meint also beides, den computergestützten Extraktionsprozess als auch den eigentlichen intellektuellen Entdeckungsvorgang.

Aus den in Literaturdatenbanken wie *Medline*, *Scisearch* u.a. ermittelten Dokumenten zu einem Themengebiet lassen sich sowohl dessen (*bibliometrische*) *Infrastruktur* (Autoren, Organisationen, Zeitschriften) als auch seine *kognitiven Strukturen* (Teilthemen, Trends, Schwerpunkte, zeitliche thematische Veränderungen) extrahieren. Hier ist besonders die *Ko-Wort-Analyse* zu erwähnen [3, 4], die für jedes (Schlag-) Wort die Häufigkeiten des gemeinsamen Auftretens mit den anderen im Literaturset vorkommenden (Schlag-) Wörtern feststellt, diese Ko-Wort-Frequenzen als Grundlage für die Gruppierung der (Schlag-) Worte (Clusterbildung) nimmt und auf diese Weise zu einer Darstellung der „semantischen Struktur“ der gegebenen Literatur gelangt. Natürlich sollten die so zustande gekommenen „Literatur-Kartierungen“ ggf. von fachlichen Experten auf ihre Gültigkeit überprüft werden.

Übersichten zu KDD, zur Wissensrepräsentation und -aufdeckung durch

Informations-Visualisierung sowie zur Ko-Wort-Analyse als Mittel der Wissensaufdeckung finden sich in [5 - 7].

Eine spezielle Methode zur Plausibilitätsprüfung von Hypothesen bzw. zur Generierung von Hypothesen wurde von Swanson [8 - 14] entwickelt: betrachtet man zwei Literaturmengen, die bezüglich ihrer Hauptthematik disparat sind, also keine in beiden Mengen vorkommenden Dokumente enthalten, so kann man ggf. durch einfachen Vergleich der Titel zu gemeinsamen, für beide Thematiken relevanten intermediären Konzepten gelangen, die Hinweise auf eine mögliche reale Verbindung der beiden „Literaturpartner“ enthalten und zur Formulierung einer durch experimentelle und klinische Forschung überprüfbarer Hypothese führen können. Swanson spricht hier von „*complementary but disparate literatures*“ [13], „*implicit, unnoticed connections*“ [11], „*undiscovered public knowledge*“ [8] oder „*logically-related noninteractive medical literatures*“ [10].

Das Vergleichen zweier Literaturmengen setzt natürlich eine wie auch immer geartete Hypothesen-Vorformulierung voraus. Für

den Fall, dass eine Hypothese erst noch gesucht wird, beschreibt Swanson einen Weg, der von einer Ausgangs-Literatur über die Eruierung möglicher interessanter intermediärer Konzepte zu komplementären (aber disparaten) Literaturmengen führen kann [13].

Eigene Untersuchungen

In der hier vorzustellenden präliminären Studie wurde das von Swanson 1986 vorgestellte Beispiel *Fish Oils / Raynaud's Disease* [8] verwendet. Online-Recherchen wurden bei DIMDI oder in PubMed durchgeführt. Die weitere Prozessierung der Retrieval-Ergebnisse (Titel-Vergleiche, Ko-Wort-Analysen) erfolgte mittels selbst erstellter perl-Skripts.

In einer in mehreren Datenbanken durchgeführten umfassenden Recherche bei DIMDI wurde festgestellt, dass es - wie von Swanson angegeben [8]- im Zeitraum bis 1985 tatsächlich keine Dokumente gibt, die sowohl den Terminus „Raynaud“ oder „Raynaud's“ als auch einen der zum Thema „Fish Oils“ gehörenden Begriffe enthalten (Abb. 1). Für die weiteren Untersuchungen wurden die Retrieval-Ergebnisse von in

```
C= 1 4677764 SELECT ME66;EM74;BA70;IS74
S= 2.00 10150 FIND CT D FISH OIL#
3.00 2062 FIND CT=ESKIMO#
4.00 28418 FIND FISH OIL#;COD LIVER OIL#;SALMON OIL#;
MENHADEN OIL#;ESKIMO#;EICOSAPENTAENOIC ACID#
5.00 30092 FIND 2;3;4
6.00 4643 FIND 5 AND PY<=1985

7.00 8439 FIND CT D RAYNAUD?
8.00 15356 FIND RAYNAUD?
9.00 15618 FIND 7;8
10.00 6244 FIND 9 AND PY<=1985

11.00 0 FIND 6 AND 10
```

Abb.1.
Rechercheprofil und Schnittmenge der Themen „Fish Oils“ und „Raynaud's Disease“.

Datenbanken: Medline (ME66), EMBASE (EM74), BIOSIS (BA70), SCISEARCH (IS74).Host:DIMDI

Abb.2. PubMed-Rechercheprofile zu den Themen „Fish Oils“ und „Raynaud's Disease“ (Vorkommen im Titel) sowie „Blood Viscosity“ (Vorkommen im Titel und als MeSH). Zeitraum 1966 bis 1985.

Abb.3.
Titel aus den beiden disparaten Literaturmengen zu „Fish-Oils“ und „Raynaud's Disease“ mit der gemeinsamen Phrase „Blood Viscosity“.

F: Fish Oils, R: Raynaud's Disease

Abb.4.
Ko-Wort-Clusteranalyse der Literatur zum Thema „Blood Viscosity“ aus PubMed (1966-1985).

Im nächsten Schritt wurde eine weitere PubMed-Recherche nach Dokumenten mit „Blood Viscosity“ im Titel bzw. im MeSH-Feld (Zeitraum bis 1985) durchgeführt. 3653 Dokumente wurden gefunden (Abb. 2 c) und einer Ko-Wort-Analyse unterworfen. Dafür wurden die Medical Subject Headings (MeSH), mit denen die Dokumente verschlagwortet sind, herangezogen. Es wurden allein Main Headings berücksichtigt. Sie wurden pro Dokument nur einmal gezählt, auch wenn sie in verschiedenen Heading/Subheading-Kombinationen vorkommen, d.h. für die Häufigkeit eines MeSH-Terms in einem Dokument wurde eine 0 (kein Vorkommen) oder eine 1 (Vorkommen) vergeben. Als Mass für die Stärke der Bindung zwischen zwei MeSH-Terms wurde der Equivalence-Index [4] berechnet: $E_{ij} = (C_{ij})^2 / C_i * C_j$, wobei C_{ij} gleich der Anzahl der Dokumente ist, in denen die Terme i und j gemeinsam vorkommen; C_i bzw. C_j ist gleich der Anzahl der Dokumente, in denen der Term i bzw. Term j vorkommt. E_{ij} kann also nur Werte von 0 bis 1 annehmen.

Term nicht auftritt. Für die anschliessende Zusammenfassung von MeSH-Terms in Gruppen (Clustering) wurden die Term-Paare nach absteigender Stärke sortiert.

Die beiden Terme im Term-Paar mit dem höchsten Equivalence-Index wurden die ersten Mitglieder des ersten Clusters. Aus den restlichen Term-Paaren wurden die herausgesucht, in denen einer der beiden als erste geclusterten Terme als Partner vorkam. Diese wurden wiederum nach absteigender Stärke sortiert; das Paar mit dem höchsten Equivalence-Index wurde dem ersten Cluster hinzugefügt, das nunmehr drei verschiedene Terme beinhaltete. Für die nächste Runde wurden dann alle Paare herausgefiltert, in denen einer der drei bereits geclusterten Terme als Partner vorkam usw. Die Clustergrösse wurde auf 20 verschiedene Terme beschränkt. Wenn also ein Cluster aufgefüllt war, wurde ein neues Cluster mit dem Term-Paar begonnen, das von allen noch vorhandenen Term-Paaren den höchsten Equivalence-Index besass. Ein neues Cluster wurde ebenfalls dann begonnen, wenn während einer Cluster-Runde alle Term-Paare mit Beteiligung der geclusterten Terms verbraucht waren. Ein beliebiges Cluster kann also zwischen 1 und 20 MeSH-Terme enthalten. Für die graphische Darstellung (s.Abb.4) wurden nur Cluster mit mindestens 4 MeSH-Terms berücksichtigt.

In einem weiteren Schritt wurde mittels der Equivalence-Indices die Verbundenheit der Begriffe innerhalb eines Clusters (Density) sowie die Stärke der Verbindungen der einzelnen Cluster zu anderen Clustern (Centrality) ermittelt. Diese Werte dienen als Koordinaten für die graphische Darstellung der Cluster (Abszisse: Centrality, Ordinate: Density; s. Abb. 4) [4].

Abb. 4 zeigt die „Blood Viscosity“-Cluster. Schnittpunkt der beiden Achsen ist der Median der Centrality-Werte und der Median der Density-Werte. Interessant ist die Lage der beiden Cluster, die den Raynaud-Term „Raynaud's Disease“ (Cluster 15) und den Fish-Oil-Term „5,8,11,14,17 - Eicosapentaenoic Acid“ (Cluster 9) enthalten: sie liegen dicht beieinander. Mit Hilfe der Ko-Wort-Analyse ist es also (in diesem Falle!) gelungen, disparate Thematiken in unmittelbare Nachbarschaft zu bringen.

Schlussfolgerung

Es mag zu früh sein, um aus diesem einen Beispiel zu folgern, dass sich disparate, aber komplementäre Themen generell mittels Ko-Wort-Techniken in benachbarte Positionen manövrieren lassen. Immerhin scheint es viel

Abb. 2 a Raynaud's Disease
Search RAYNAUD* Field: Title Word, Limits: Publication Date from 1966 to 1985
801

Abb. 2 b Fish Oils
Search FISH OIL OR FISH OILS OR COD LIVER OIL OR COD LIVER OILS OR SALMON OIL OR SALMON OILS OR MENHADEN OIL OR MENHADEN OILS OR EICOSAPENTAENOIC ACID OR EICOSAPENTAENOIC ACIDS
Field: Title Word, Limits: Publication Date from 1966 to 1985
248

Abb. 2 c Blood Viscosity
Search BLOOD VISCOSITY[title] OR BLOOD VISCOSITY[mh]
Limits: Publication Date from 1966 to 1985
3653

PubMed durchgeführten Recherchen nach Dokumenten, die die Begriffe aus Abb.1 im Titel führen, herangezogen (Abb. 2 a, 2 b). Durch Titelvergleich wurden u.a. die in Abb. 3 dargestellten Raynaud- und Fish-Oil-Titel gefunden, die auf das intermediäre Konzept „Blood Viscosity“ verweisen.

$E_{ij} = 0$ bedeutet, Term i und Term j kommen keinmal gemeinsam in einem Dokument vor. $E_{ij} = 1$ bedeutet, dass Term i und Term j nur gemeinsam vorkommen. Zwischen 0 und 1 liegende Werte von E_{ij} bedeuten, dass mindestens einer der beiden Terme auch in Dokumenten vorkommt, in denen der andere

F Beneficial effect of fish oil on blood viscosity in peripheral vascular disease.
Reduction in blood viscosity by eicosapentaenoic acid.

R Raynaud's phenomenon and blood viscosity.
Local increase of blood viscosity during cold-induced Raynaud's phenomenon.

Abbildung 3

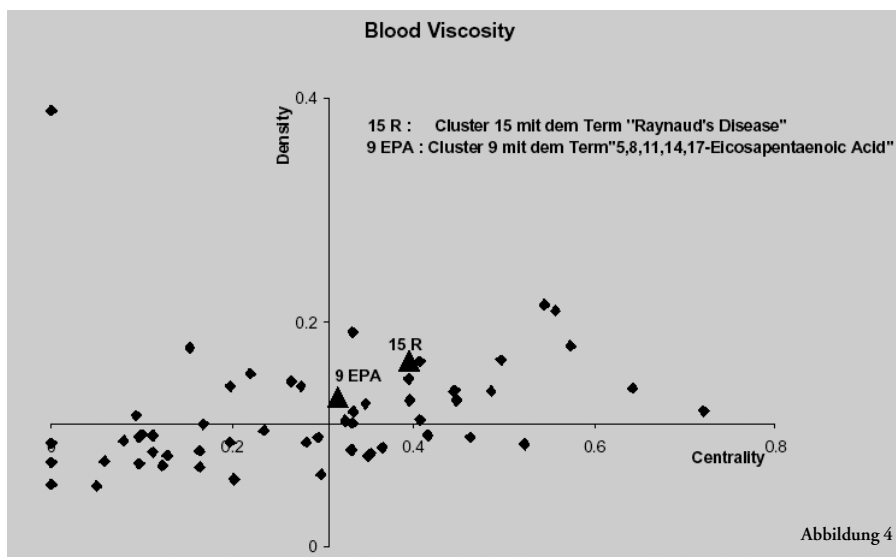


Abbildung 4

versprechend, diese Phänomene weiter zu erforschen. Im übrigen sind wissenschaftliche Bibliotheken mit ihren Datenbank-Kenntnissen und ihrem Retrieval-Know-how geradezu prädestiniert dafür, als Mediatoren zwischen Informations- und Fachwissenschaft zu fungieren. Dabei ist sicherlich von Vorteil, die Methoden und Techniken auch selber anwenden zu können; dies ist erst recht notwendig, wenn man Informationssysteme konzipieren und realisieren will, die „mehr“ aus Literatur machen und ggf. Hinweise auf vielversprechende neue Hypothesen generieren können.

1. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. Knowledge discovery and data mining: towards a unifying framework. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, Oregon, August 2-4, 1996, AAAI-Press.
2. Kostoff, R.N. and DeMarco, R.A. Extracting information from the literature by text mining. *Analytical Chemistry* 73 (13), 2001, 370A-378A.
3. Callon, M., Law, J. and Rip, A. (eds.) *Mapping the dynamics of science and technology*. 1986, Macmillan Press Ltd. London.
4. Callon, M., Courtial, J.P. and Laville, F. Co-Word analysis as a tool for describing the network of interactions between basic and technological research - the case of polymer chemistry. *Scientometrics* 22 (1), 1991, 155-205.
5. Trybula, W.J. Data mining and knowledge discovery. *Annual Review of Information Science and Technology* 32, 1997, 197-229.

6. White, H.D. and McCain, K.W. Visualization of literatures. *Annual Review of Information Science and Technology* 32, 1997, 99-168.
7. He, Q. Knowledge discovery through Co-Word analysis. *Library Trends* 48 (1), 1999, 133-195.
8. Swanson, D.R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine* 30 (1), 1986, 7-18.
9. Swanson, D.R. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine* 31 (4), 1988, 526-557.
10. Swanson, D.R. Online search for logically-related noninteractive medical literatures: a systematic trial-and-error strategy. *Journal of the American Society for Information Science* 40 (5), 1989, 356-358.
11. Swanson, D.R. A second example of mutually isolated medical literatures related by implicit, unnoticed connections. *Journal of the American Society for Information Science* 40 (6), 1989, 432-435.
12. Swanson, D.R. Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association* 78 (1), 1990, 29-37.
13. Swanson, D.R. and Smalheiser, N.R. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence* 91 (2), 1997, 183-203.
14. Swanson, D.R. and Smalheiser, N.R. Implicit text linkages between Medline records: using Arrowsmith as an aid to scientific discovery. *Library Trends* 48 (1), 1999, 48-59.

Johannes Stegmann
 Freie Universität Berlin
 Medizinische Bibliothek im Universitätsklinikum Benjamin Franklin
 Hindenburgdamm 30
 12203 Berlin
 email:b4johannes.stegmann@medizin.fu-berlin.de

Publikationen:

Tips and tricks for website managers / ed. by Mark Kerr. - London: Aslib-IMI, 2001. - 170 S. - ISBN 0-85142-439-2 (19.99 £)

Claudia Lascar ; Loren D. Mendelsohn: An analysis of journal use by structural biologists with applications for journal collection development decisions. - In: *College & Research Libraries* 62 (2001) 5, S. 422 - 433.

David Nicholas, Peter Williams, Paul Huntington: Health information kiosk use in health organisations: the views of the health professionals. - In: *Aslib Proceedings* 53 (2001) 9, S. 368 - 386.

[health kiosk = touch screen information kiosk in medical locations / Untersuchung der Akzeptanz bei Patienten, Kindern, Schwestern, Ärzten und Verwaltung]

Consumer Health: A guide to Internet information resources / by Cecilia Durkin. (\$ 68) - mit einer CD-ROM, die die Links

zu den besprochenen Websites enthält. www.mlanet.org/publications (A. Fulda)

