Factor Analytic Approach to Transitive Text Mining Using PubMed Keywords[1]

Johannes Stegmann* and Guenter Grohmann**

*johannes.stegmann@charite.de
Medical Library, Charité - University Medicine Berlin,
Campus Benjamin Franklin, 12203 Berlin, Germany

**guenter.grohmann@charite.de
Institute of Medical Informatics, Biometry and Epidemiology, Charité - University Medicine Berlin,
Campus Benjamin Franklin, 12203 Berlin, Germany

Abstract

Matrix decomposition methods were applied to examples of non-interactive literature sets sharing implicit relations. Document-by-term matrices were created from downloaded PubMed literature sets, the terms being the Medical Subject Headings (MeSH descriptors) assigned to the documents. The loadings of the factors derived from singular value or eigenvalue matrix decomposition were sorted according to absolute values and subsequently inspected for positions of terms relevant to the discovery of hidden connections.

It was found that only a small number of factors had to be screened to find key terms in close neighbourhood, being separated by a small number of terms only. It is concluded that in literature-based discovery processes the decomposition methods combined with human inspection of the created factors may play an important role, provided MeSH descriptors are analysed.

Introduction

Transitive text mining tries to establish meaningful links between the main concepts of non-overlapping literature sets. The basic notion of this kind of literatures as "complementary but disjoint" (CBD) as well as several examples of literature-based "discoveries" from the medical literature have been published by Swanson (e.g. Swanson, 1986; Swanson, 1988, Swanson, 1991). Different methods have been applied to detect possibly useful links between "non-interactive medical literatures" (Swanson, 1989), either using words taken from titles and abstracts (Swanson & Smalheiser, 1999; Gordon & Lindsay, 1996; Gordon & Dumais, 1998; Weeber et al, 2001) or analysing the Medical Subject Headings (MeSH) assigned to the indexed documents (Stegmann & Grohmann, 2003; Srinivasan, 2004). All types of approaches try to find intermediate terms and concepts in order to retrieve literature which contains both, the otherwise non-interactive "source" and "target" terms. The MeSH-based approach developed by us extracts the MeSH descriptors from the downloaded documents and tries to locate relevant source, intermediate and target terms in two-dimensional cluster diagrams of the MeSH terms according to some positional and/or numerical characteristics (Stegmann & Grohmann, 2003). Dealing, however, with large literature sets and high numbers of terms hampers cluster localisation and screening. Therefore, we are interested in other methods to find channels from source to target concepts. The investigation presented here is a variation of an approach which involves matrix factorization methods. A general form of factor analysis is singular value decomposition (svd). Svd has been introduced as a tool for "latent semantic analysis" (LSA) by Deerwester et al. (1990) and was applied by Gordon & Dumais (1998) to establish the chain from Raynaud's Disease to Fish Oil by dimensional reduction of the space spanned by documents and text words. Eigenvalue decomposition (evd) is closely related to svd and was applied in the present study. The idea is to find within the first few factors of the resulting factor matrix interesting intermediate and target terms in close neigbourhood to source terms.

The literature sets tested represent two tracks of CBD literatures: the classical "Raynaud's Disease / Fish Oil" example (Swanson, 1986) and the "Cardiac Hypertrophy / Chlorpromazine" example recently described by Wren et al. (2004).

[1] Our work is currently supported by the Deutsche Forschungsgemeinschaft, grant no. LIS 4 - 542 81

Proceedings of ISSI 2005, pages 280–283.
Edited by P. Ingwersen and B. Larsen.

Methods

PubMed title searches (restricted to publication years 1966 to 1985) for "Raynaud's Disease" and its related themes were performed between June and November 2004 as described (Stegmann & Grohmann, 2003). Literatures on "Cardiac Hypertrophy", "Chlorpromazine" and "Norepinephrine" were retrieved by appropriate title searches (not shown) from PubMed in November 2004. MeSH descriptors were extracted from each downloaded document set by means of homemade PERL or JAVA scripts. Binary document-by-term matrices were produced with document numbers as rows, MeSH terms as columns, and "0" or "1" as values of the matrix cells. These matrices - being of sizes between 287 and 8858 rows and 332 and 3509 columns (not shown) - were left-multiplied with the respective transposed and subjected to evd, using the software package R (R Development Core Team, 2004). The resulting factorised matrix contains as rows the MeSH descriptors and as columns the factors constituted by the MeSH terms. The matrix elements - the factor loadings - were converted to absolute values (Kostoff et al, 2004), sorted column by column and inspected to find the positions of source, intermediate and target terms.

Results

Table 1 summarises the results found for the "Raynaud's Disease / Fish Oil" track. In the cases of source ("Raynaud's Disease") and target ("Fish Oil") literature one has to screen only few columns of the factor matrices to find source or target terms close to intermediate terms. More factors have to be screened in the cases of the intermediate literatures; however, in each case the number of columns which must be screened until relevant terms are found near each other is well below 3%. In addition, it is obviously sufficient to inspect a number of nearest terms above and/or below the guiding terms corresponding to less than 5% of all terms (Table 1).

The potential target terms "Nitric Oxide" and "Arginine" (Stegmann & Grohmann, 2003) are also found in the neighbourhood of "Raynaud's Disease" within reasonable numbers of factors/terms in the "Platelet Aggregation" factor matrix (Table 1).

Table 1. Raynaud's Disease – Fish Oil literature track: positions and distances of source, intermediate and target terms after factorization. Abbreviations: src: source, irm: intermediate, tar: target, RD: Raynaud's Disease, BV: Blood Viscosity, FO: Fish Oils, EPA:Eicosapentaenoic acid, PA: Platelet Aggregation, ARG: Arginine, NO: Nitric Oxide

Literature theme, type, size (no. documents x no. of MeSH terms)	Guiding term			Term screened for in neighbourhood of guiding term			Distance (no. of terms in between)
	term	factor	position	type	term	position	
Raynaud's Disease,	RD	2	1	irm	BV	20	18
Source (801 x 464)	RD	5	53	irm	BV	56	2
	RD	10	73	irm	PA	61	11
Fish Oil,	FO	1	6	irm	PA	15	8
Target (287 x 332)	EPA	2	2	irm	PA	11	8
	FO	8	39	irm	BV	42	2
Blood Viscosity,	RD	10	64	tar	EPA	73	8
intermediate (502 x 393)	RD	23	35	tar	EPA	38	2
Platelet Aggregation,	RD	12	447	tar	EPA	471	23
intermediate	RD	21	775	tar	ARG	754	20
(2636 x 1532)	RD	32	776	tar	FO	787	11
	RD	37	614	tar	NO	617	2

Next, we applied the factorisation method to the literatures of the "Cardiac Hypertrophy / Chlorpromazine" track. The results are shown in Table 2. The theme "Cardiac Hypertrophy" is expressed by several MeSH terms. The drug theme is represented by the term "Chlorpromazine". The possible intermediate term "Norepinephrine" (a widespread neurotransmitter) which was mentioned by Wren et al. (2004) as one of the relatively high ranking relationships shared between "Cardiac

Hypertrophy" and Chlorpromazine" is found in close neighbourhood of the source terms within the first 10 factors of the cardiac hypertrophy literature set. Equally, the term "Norepinephrine" is found close to "Chlorpromazine" within the first few factors derived from the target literature on "Chlorpromazine". In the case of the intermediate "Norepinephrine" literature the terms "Cardiomegaly" and "Chlorpromazine" have a distance of 27 only (less than 1% of the maximally possible distance) already in the first factor, and in factor 18 the distance between source and target terms is yet considerably smaller (Table 2).

Table 2. Cardiac Hypertrophy – Chlorpromazine literature track: positions and distances of source, intermediate and target terms after factorization. Abbreviations: CM: Cardiomegaly, NE: Norepinephrine, CP: Chlorpromazine, HL: Hypertrophy, Left Ventricular, HR: Hypertrophy, Right VentricularFurther details: see Table 1.

Literature theme, type, size (no. documents x no. of MeSH terms)	Guiding term			Term screened for in neighbourhood of guiding term			Distance (no. of terms in between)
	term	factor	position	type	term	position	
Cardiac Hypertrophy,,	CM	8	31	irm	NE	41	9
Source (5955 x 2146)	HR	9	48	irm	NE	50	1
Chlorpromazine,	CP	2	25	irm	NE	45	19
Target (3532 x 2396)	CP	11	29	irm	NE	32	2
Norepinephrine,	CM	1	337	tar	CP	365	27
intermediate	HL	18	2884	tar	CP	2877	6
(8858 x 3509)							

Discussion

Singular value decomposition was used by Gordon & Dumais (1998) as a tool to reduce the dimensions of the knowledge space spanned by documents and title or abstract words or phrases they analysed. They re-built the decomposed matrix on a considerably lower dimensional level and determined the similarity between terms using the cosine measure. They found intermediate concepts near to the source concept "Raynaud's Disease"; however, analysing intermediate literature, the target concepts "Eicosapentaenoic acid" and "Fish Oil" were located far apart from the the source concept, disabling a quick successful screen of the terms.

Our investigation presented here uses the descriptors assigned to PubMed (Medline) documents, thus accomplishing a considerable reduction in dimensionality. Although the information content of the documents certainly cannot be fully described by the MeSH terms (Kostoff et al., 2004), it has been shown that MeSH terms are well suited to support the discovery process (Stegmann & Grohmann, 2003; Srinivasan, 2004). In the present study, the factors created by eigenvalue decomposition were directly analysed, omitting the further steps performed by Gordon & Dumais (1998). Relatively easy and fast screening of only a few factors found relevant terms in close neighbourhood. The method described may complement our two-dimensional cluster map approach described earlier (Stegmann & Grohmann, 2003), especially in analyses dealing with large literature sets and high numbers of terms and clusters.

The investigation presented here uses the decomposition techniques and the subsequent manipulations simply as a heuristic means to detect terms relevant to the discovery of hidden connections and does not try to uncover any mechanism possibly underlying the collocation processes, work that certainly should be done in the future. In addition, it should be investigated whether relevant terms taken from titles and abstracts can be collocated by the procedure described in this presentation.

Conclusion

A method was described which uses factorised document-by-MeSH-term matrices in order to find - by quick and easy manual screening - in close neighbourhood terms being of interest in the process of linking disparate but complementary literatures. The method supplements the already existing approaches to literature-based hypothesis generation.

References

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R. (1990), Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391-407.

Gordon, M.D. & Lindsay, R. K. (1996). Toward discovery support systems: a replication, re-examination and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *Journal of the American Society for Information Science*, 47, 116-128.

Gordon, M. D. & Dumais, S. (1998). Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science*, 49, 674-685.

Kostoff, R.N., Block, J.A., Stump, J. A. & Pfeil, K.M. (2004). Information content in Medline record fields. *International Journal of Medical Informatics*, 73, 515-527.

R Development Core Team (2004), *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Srinivasan, P. (2004). Text Mining: generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology,* 55, 396-413.

Stegmann, J. & Grohmann, G. (2003). Hypothesis generation guided by co-word clustering. *Scientometrics,* 56 , 111-135.

Swanson, D.R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30, 7-18.

Swanson, D.R. (1988). Migraine and magnesium: eleven neglected connections. Perspectives in Biology and Medicine, 31, 526-557.

Swanson, D.R. (1989). Online search for logically-related noninteractive medical literatures: a systematic trial-and-error strategy. *Journal of the American Society for Information Science*, 40, 356-358.

Swanson, D. R. (1991). Complementary Structures in disjoint literatures. In A. Bookstein, Y. Chiaramella, G. Salton, V. V. Raghavan (Eds.), *SIGIR '91: Proceedings of the 14th annual international ACM / SIGIR Conference on Research and Development in Information Retrieval* (pp 280-289). New York: ACM.

Swanson, D.R. & Smalheiser, N.R. (1999). Implicit text linkages between Medline records: using Arrowsmith as an aid to scientific discovery. *Library Trends*, 48, 48-59.

Weeber, M., Klein, H., de Jong-van den Berg, L. T. W. & Vos, R. (2001). Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine- magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52, 548-557.

Wren, J.D., Bekeredjian, R., Stewart, J.A., Shohet, R.V. & Gamer, H.R. (2004). Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, 20, 389-398.