Chute, C. (1991). The Classification of Medical Events Using Latent Semantic Analysis. 2nd ASIS SIG/CR Classification Research Workshop, 41-48.
doi:10.7152/acro.v2i1.12546

The Classification of Medical Events Using Latent Semantic Analysis

C. G. Chute, M.D., Dr.P.H.
Section of Medical Information Resources,
Mayo Clinic
Rochester, MN 55905

Clinical information is dominated by natural language representation of data and knowledge. To bring quantitative methods to bear in the empiric analysis of clinical episodes, they must be classified into reasonably homogenous categories that sustain inference and generalization. A tangible, if trivial, example of a classification requirement is the retrieval of patient cases relevant to the testing of a clinical hypothesis, so that they can be further scrutinized. Reliance on text word retrieval alone, drawn from natural language summaries, is fraught with contextual ambiguity and defeated by an expressively rich sub-language.

MEDICAL CLASSIFICATION

The recognition of classification as a problem in disease description dates to the 17th century London Bills of Mortality compiled by John of Graunt[1]. This work is heralded as the advent of modern epidemiologic research, but the recognition that efficient medical outcomes research is highly dependent on classification principles is only recently emerging. Historically, myriads of classification systems have evolved, ranging from the monolithic heritage of the Bertillon classification[2], still with us as the International Classification of Diseases (ICD)[3], to the multi-axial hierarchies of the Systematized Nomenclature of Medicine (SNOMED)[4].

Medical researchers today confront a “Tower of Babel” in clinical data repositories. Specialized reimbursement codes, procedure and disease rubrics, subspecialty systems, and scores of emerging “standards” beleaguer the investigator trying to link patient conditions, events, or outcomes. Such linkage is the very essence of health services research, epidemiology, and biostatistics. Absence of coherent classifications, or alternatively a super-classification, greatly hampers progress in these fields.

The National Library of Medicine (NLM), recognizing this growing problem, initiated work on the Unified Medical Language System (UMLS) in 1986 [5][6]. Central to this resource is a metathesaurus of composite biomedical terminologies, coordinated as a relational database of medical “main concepts”[7]. We established that the medical records relative to standard patient nomenclatures[8]. It also forms the basis of the classification and information retrieval research presented here.

A MEDICAL CONCEPT

Ideally, researchers and insurers alike aspire to have clinical events classified against fundamental medical concepts. Multi-axial systems such as SNOMED[3] attempt to enumerate “atomic” medical concepts, and then “assemble” composite or complex clinical descriptions from these parts.

The question of what constitutes a medical concept is essentially undefined. Most classifications regard “lung cancer” as a basic concept, but it could defensibly derive from the combination of

ISSN: 2324-9773

Chute, C. (1991). The Classification of Medical Events Using Latent Semantic Analysis. 2nd ASIS SIG/CR Classification Research Workshop, 41-48.
doi:10.7152/acro.v2i1.12546

"lung" and "cancer." If allowance is made for the biologic reality that bronchogenic tumors are unique and consistent unto themselves, then where does one draw the line around a concept. Should we regard "Stage III adenosquamous bronchogenic carcinoma of the left upper lobe with C-myconcogene production" as another discrete concept? This problem of granularity is not new, but is especially acute in medical data classifications.

The attribute/value nature of medical data also frustrates classification efforts. Cancer stages and grades, lab values and deviations, procedure results or strip chart tracings represent data that defies definitive classification. The concept of "anemia" exemplifies this, which typically specifies a threshold value for a serum hemoglobin level, but this threshold is conventionally higher in men than women. Further, what do we do with profound anemia, greatly below the threshold. This introduces the severity of a state or condition, raising the whole question of concept modifier.

The classification of a "history of myocardial infarction" in a healthy person is quite distinct from one happening in the present. Modifier axis of time, severity, anatomic location, extent, qualitative and quantitative aspects, probability, and spatial direction represent but a few of the plethora of combinatorials that can greatly distort the sense of concept classification.

LATENT SEMANTIC INDEXING (LSI))

The application of singular value decomposition (a computationally intense but reliable matrix factorization) to a "factor analysis" sort of information retrieval has been explored and described by Deerwester, et al.[9]. Document retrieval is undertaken by exploiting the crude semantic content of a word and term co-occurrence among text documents. An information matrix is constructed with all words encountered defining the rows, and the target documents themselves defining the columns.

Evans has proposed the designation of medical concepts as the columns in place of simple documents[10]. This has the intriguing consequence that pseudodocuments can be fashioned which detail related words and terms that should be semantically "attached" in creating an information matrix. The problems of word level lexical variation, synonyms, inflection, or noun/adverb forms can be partially addressed in such a pseudodocument. The extent to which the UMLS can serve as a pseudodocument for LSI has been initially considered[11], and is illustrated by example in Table 1.

ISSN: 2324-9773

Table 1. Pseudodocument Example from UMLS with Terms Bearing Semantic Relationship to Myocardial Infarction

MC:Myocardial Infarction
AN:Diseases (Non MeSH)
AN:Cardiovascular Diseases
AN:Heart Diseases
AN:Coronary Disease
AS:<Heart>/<Infarction>
BT:Infarction
BT:Infarctions
CH:Shock, Cardiogenic
LV:Infarct; myocardial
LV:Infarct; myocardium
LV:Infarction, Myocardial
LV:Infarction; myocardial
LV:Infarction; myocardium
LV:Infarctions, Myocardial
LV:Myocardial infarct
LV:MYOCARDIAL INFARCT NOS
LV:Myocardial Infarctions
NT:ACUTE MYOCARDIAL INFARCTION
RT:MYOCARDIAL INFARCT NEC
SI:Angina Pectoris
SI:Coronary Aneurysm
SI:Coronary Arteriosclerosis
SI:Coronary Thrombosis
SI:Coronary Vasospasm
SY:Heart Attack

Key:
MC main concept; AN ancestor
AS associated; BT broader
CH child; LV Lexical variant
NT narrower; RT related
SI sibling; SY synonym

Consideration of the underlying classification issues invite further questions however, most notably that of partial matches over concept vectors. Expansion of the LSI process, and its outcome, will facilitate such consideration.

The method outlined by Deerwester[9] exemplifies the LSI process. Briefly, given an input matrix X, the process of the singular value decomposition (SVD), X=UoSoVo, yields three outputs: the singular values So, Uo (concept x word dimensions), and Vo (a square matrix, with rank equal to the smaller dimension of the input matrix, typically the number of concepts). Each array has one dimension truncated to N, which compresses the semantic space and reduces the computational demand of applying the decomposition. The truncation yields an approximation X~Xq=USV where U is number of canonical words x N, V matrix is number of concepts in input file x N, and S includes the N most significant singular values. U, V and S are used in mapping an inquiry phrase to the term space. Inquiry phrases or medical text to be classified are canonicalized and represented as a vector Xq similar to a column of X. Xq is transformed into the new vector space: Vq=XqU(1/S). Finally, the distance between the inquiry and each term is measured by the cosine theta:

$$\text{cosine theta} = \frac{Vq.\ Vi}{|Vq| \times |Vi|}$$

where Vq is the transformed inquiry vector, Vi is a column of the V matrix; "." = dot product, | | = Euclidean norm.

The SVD can be considered to "rank" the N most important dimensions of the concepts (columns) and canonical terms (rows), derived from the input matrix. Inquiries and classification practices invoke only these dimensions, by using the truncated solution matrices U and V.

VECTOR INQUIRIES

It can be seen that the cosine theta which results from this process is a vector. Thus, a cosine value can be computed from an inquiry phrase or text requiring classification for every concept (column) in the initial information matrix. This implies that partial matches are adroitly handled by multiple dominant cosine values in the cosine vector. Terms that are semantically "opposite" tend to create negative cosines.

Text or phrases that require classification can be processed using LSI to yield a vector of cosines representing degree of match to all concepts defined in the matrix. The most significant subset, or conceivably the entire vector, can then be stored as the index for that text. Repository indices thereby constitute an array of these vectors, one for each record. Inquiry against a repository involves the creation of a similar vector using LSI for the inquiry text, then computing the euclidean distance of the inquiry vector over each row in the repository array. A composite match score is created across multiple classification concepts at one time for each repository record.

COMPLEX HIERARCHIES

The UMLS shares hierarchical characteristics with many classification systems. As such, ancestral and child relationships may be desirable to express in the LSI pseudodocuments. The notion of invoking the imaginary component of complex numbers [a number pair consisting of real and imaginary (recall the square root of -1) parts] to contain this hierarchical access is appealing, and consistent with methodologies used to create the information matrix. This raises the related question of establishing the real number weights, which should be "less" for distant hierarchical relations than for full synonyms or the main concept words themselves.

The assignment of hierarchical direction to an imaginary number component invites conflict when a word appears in ancestral contexts as well as the main concept (e.g., "heart" disease is ancestral

Chute, C. (1991). The Classification of Medical Events Using Latent Semantic Analysis. 2nd ASIS SIG/CR Classification Research Workshop, 41-48.
doi:10.7152/acro.v2i1.12546

to "heart" attack). What imaginary value should "heart" take in the myocardial infarction pseudodocument? Our preliminary solution computes the weighted average of the imaginary component contenders, using the real number component as the weights. Initially, we assigned the real number weight of 0.5 to all hierarchically distant terms, and 1.0 to closely related terms such as synonyms.

PRELIMINARY ANALYSIS

The initial CD ROM release (September 1990) of the UMLS Metathesaurus formed the basis of our evaluations. Detailed description of our methodologies appears elsewhere[12]. Briefly, we extracted all 2,580 main concepts with a semantic type of "disease or syndrome." These constituted the anchors for 2,580 pseudodocuments. Synonyms, related terms, lexical variants, and a host of hierarchical elements are relationally linked to main concepts in the UMLS. We followed the links and filled out our pseudodocuments.

An information matrix was created having 2,580 columns corresponding to these concepts. Reducing every word in every pseudodocument to canonical form using the morph tool[13], yielded 5,434 unique root words, defining matrix rows. We solved a smaller matrix of 3,803 word stem/rows by ignoring all distant hierarchical relationships. Even so, this "smaller" SVD consumed 6,000 seconds on a Cray-2. The largest cosine findings for each of three inquiry vectors are shown in Figure 1. A tiny demonstration matrix was prepared to explore the complex number representation of hierarchical elements. Only ten concepts are defined in this matrix, implying only ten pseudodocuments and corresponding columns. Output for this matrix appears in Figure 2, including the complex cosine values.

Figure 1.
Top Level Matches for Real Number Decomposition of Large Matrix

Input Phrase	Matched Phrase	Cosine Deviation
carcinoma of the lung	CARCINOMA OF LUNG	1.00
	Carcinoma, Non-Small Cell Lung	1.00
	Lung Neoplasms	1.00
	Pleural Neoplasms	1.00
	Bronchial Neoplasms	0.85
cerebral ischemia	Cerebral Ischemia	0.80
	Cerebral Infarction	0.64
	Encephalomalacia	0.64
	Brain Damage, Chronic	0.60
	Cerebral Ischemia, Transient	0.60
myocardial infarction	Myocardial Infarction	0.78
	Myocardial Reperfusion Injury	0.65
	Myocardial Diseases	0.59
	Angina, Unstable	0.48
	Cerebral Infarction	0.25

ISSN: 2324-9773

Figure 2.
Complete LSI Process Output for 10-Concept Matrix

Input Phrase	Matched Phrase	Cosine Deviation
carcinoma of the lung	Lung Neoplasms	(0.99,-0.01)
	Coin Lesion, Pulmonary	(0.99,0.01)
	Cholecystitis	(0.42,-0.10)
	Cholelithiasis	(0.18,-0.15)
	Kidney Failure, Acute	(0.17.-0.11)
cerebral ischemia	Stroke	(0.98,-0.02)
	Cerebral Infarction	(0.91,0.00)
	Coronary Arteriosclerosis	(0.18,-0.06)
	Cholecystitis	(0.03,-0.03)
	Myocardial Infarction	(0.13,-0.11)
myocardial infarction	Myocardial Infarction	(0.93,-0.03)
	Coronary Arterioscelrosis	(0.71,-0.06)
	Cholecystitis	(0.26,-0.17)
	Cerebral Infarction	(-0.00,-0.02)
	Coin Lesion, Pulmonary	(-0.02,-0.01)

Note that the cosine of the concept deviations are complex values.

DISCUSSION

LSI suggests a new paradigm for classification of text entities. The truncation of the SVD solution matrices to a smaller number (N: 10% to 25% of the original number of concepts) implies the definition of a compressed concept space, in which only the most significant components of a factor analysis exist. These compressed components may be regarded as abstract concepts, derived from the composite semantic information contained in the pseudodocuments (hence the name: latent semantic indexing). The abstraction is implied from the "collapsing" of closely related terms across concepts into N "metaterms"; in the V matrix, analogously the closely related concepts are collapsed into N "metaconcepts" across terms in the U matrix by the SVD.

The LSI process creates an abstract semantic network from readily expressed term relationships, without the tedium and maintenance incumbent with semantic network construction. Indeed, it creates such a structure from a textual relational database off the shelf--the UMLS Metathesaurus.

Comparison of these abstract concepts to neural network nodes is beyond the scope of this paper. Superficially however, the LSI process may simulate a pre-coordinated, static neural network that cannot "learn," but can be "taught" once. The SVD solution constitutes this instruction. Parameters of initial matrix construction before decomposition might be regarded as instruction, but this is before the birth of an abstract semantic space.

Chute, C. (1991). The Classification of Medical Events Using Latent Semantic Analysis. 2nd ASIS SIG/CR Classification Research Workshop, 41-48. doi:10.7152/acro.v2i1.12546

The potential for LSI to practically impact a medical classification may be enormous. Vigorous pursuit of these evaluations seems warranted.

REFERENCES

1. Graunt J. *The Natural and Political Observations Made Upon the Bill of Mortality.* Baltimore, MD: Johns Hopkins University Press, 1939. [Reprinted from the original, London, 1662.]
2. Bertillon J. Classifications of the causes of death. Transactions of the 15th International Congress on Hygiene and Demographics. Washington, D.C., 1912:52-55.
3. International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM). Ann Arbor, MI: Commission on Professional and Hospital Activities, 1986.
4. Cote RA. Systematized Nomenclature of Medicine (SNOMED). Skokie, IL: College of American Pathologists, 1982.
5. Unified Medical Language System. National Library of Medicine News 1986;41(11):1-2,10-11.
6. Lindberg DAB, Humphreys BL. The UMLS Knowledge Sources: Tools for building better user interfaces. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care 1990;121-125.
7. Tuttle M, Sherertz D, Olson N, Erlbaum M, Sperzel D, Fuller L, Nelson S. Using Meta-1 -- the 1st version of the UMLS Metathesaurus. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care 1990;131-135.
8. Chute CG, Yang Y, Tuttle MS, Sherertz DD, Olson NE, Erlbaum MS. A preliminary evaluation of the UMLS Metathesaurus for patient record classification. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care 1990:161-165.
9. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 1990;41(6):391-407.
10. Evans DA, Handerson SK, Monarch IA, Pereiro J, Delon L, Hersh WR. Mapping vocabularies using "Latent Semantics". Technical Report No. CMU-LCL-91-1, Pittsburgh, PA: Computational Linguistics Laboratory, Carnegie Mellon University, 1991.
11. Chute CG, Yang Y, Evans DA. Latent Semantic Indexing of medical diagnoses using UMLS Semantic Structures. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care 1991, in press.
12. Yang Y, Chute CG. A schematic analysis of the Unified Medical Language System. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care 1991, in press.
13. Evans DA, Ginther-Webster K, Hart M, Lefferts RG, Monarch IA. Automatic indexing using selective NLP and First-Order Thesauri. RIAO '91, April 2-5, 1991, Autonoma University of Barcelona, Barcelona, Spain, pp. 624-644.

ISSN: 2324-9773

Chute, C. (1991). The Classification of Medical Events Using Latent Semantic Analysis. 2nd ASIS SIG/CR Classification Research Workshop, 41-48.
doi:10.7152/acro.v2i1.12546

ISSN: 2324-9773