An Evaluation of Concept Based Latent Semantic Indexing for Clinical Information Retrieval

Christopher G. Chute, M.D., Dr.P.H. and Yiming Yang, Ph.D.

Section of Medical Information Resources
Department of Health Sciences Research
Mayo Clinic
Rochester, Minnesota

ABSTRACT

Latent Semantic Indexing (LSI) of surgical case report text using ICD-9-CM procedure codes and index terms was evaluated. The precision-recall performance of this two-step matrix retrieval process was compared with the SMART Document retrieval system, surface word matching, and humanly assigned procedure codes. Human coding performed best, two-step LSI did less well than surface matching or SMART. This evaluation suggests that concept-based LSI may be compromised by its two-stage nature and its dependence upon a robust term database linked to main concepts. However, the potential elegance of partial- credit concept matching merits the continued evaluation of LSI for clinical case retrieval.

INTRODUCTION

Retrieval of patient data for outcome analysis, continuous improvement, clinical epidemiology, or practice description is becoming an increasingly important activity in health care delivery and research. Text word searches or traditional document retrieval software function poorly with dictated medical text, due to the concept-based nature of most inquiries (e.g. colonic neoplasia) and the vagaries of real word medical vocabulary which fail to "match" (e.g. large bowel cancer).

Traditionally, practical medical information retrieval has been based on the human coding of clinical text into medical classifications or nomenclatures such as ICD-9-CM [1] or SNOMED-II [2]. These systems impose a coarse conceptual framework for retrieval, but often sacrifice specificity and are expensive to routinely apply across all the dictations created in patient care.

We previously proposed Latent Semantic Indexing (LSI) as a technique to invoke an available database of synonyms and lexical variants associated with a main medical concept (the UMLS Metathesaurus) to assign medical text to a base concept [3]. Our preliminary evaluation showed this use of an off-the-shelf concept synonym collection held promise, and afforded the idea of partial credit matches of specific text to several concepts with varying association strengths or similarity. The present work expands these evaluations to larger volumes of clinical text and compares case retrieval performance to alternative information retrieval approaches.

METHODS

Clinical Data Resource. The 94,337 online surgical records of all Mayo Clinic patients during 1991 formed the information base for our retrieval experiments. To reduce the burden of human review, we restricted the database to the 6,555 cases which had an ICD-9-CM procedure code [4] between 35.00 to 39.99 (Operations on the Cardiovascular System). Because we employed a 1980 public-domain version of the ICD-9-CM and its index, we excluded 1,043 cases having codes (e.g. 36.01) added since 1980. Among the 55 data fields for each case, we selected ICD-9-CM procedure codes, physician dictated post-operative procedure titles, and the full text of the operative report. An arbitrary sequential case number was assigned to each report as an indexing key.

Three databases were created from surgical sources based on text deriving from: 1) the procedure titles only; 2) operative report only; and 3) combined procedure titles and operative report. For each text database, all words and terms were reduced to a preferred form (canonicalized) using a 97,037 element clinical lexicon edited by our research team and the `morph` component of the CLARIT system from CMU [5].

Canonical Concepts for Latent Semantics. A machine readable copy of the public domain version of the ICD-9-CM [6] formed the basis of our latent semantic data. For each of the 259 procedure codes between 35.00 and 39.99, all main entry text and indexing terms associated with the rubric were canonicalized. Pooling across these terms yielded 700 canonical words from which 178 were excluded as articles, prepositions, conjunctions, or

0195-4210/92/$5.00

informationless words as characterized in our lexicon, and 9 were mapped to synonyms. The resulting 513 words constituted the row entries for the Latent Semantic Indexing (LSI) information matrix, described previously [3]. The 259 *ICD-9-CM* codes comprised the matrix concepts as columns.

The 513 by 259 element matrix was populated by assigning a 1 to the cell where a preferred form word (row) is associated with the *ICD-9-CM* code column, and a 0 otherwise. This matrix was then reduced by singular value decomposition (SVD) [7] to yield the partial solution matrices for truncation and LSI retrieval.

The Unified Medical Language System Metathesaurus (UMLS, version 1.1) contains only 25 concepts which are related to or descended from "Surgery, Cardiovascular." We conducted a parallel process of LSI-based classification and retrieval evaluation from these 25 terms. The results reflect the intention of the present UMLS version *not* to serve as a clinical classification by showing a very poor performance of UMLS-based LSI. Because our data were exclusively clinical, we elected not to present these results based on the present UMLS version. This will form the basis of a future report when the UMLS is more clinically enriched, perhaps including all of *ICD-9-CM*.

Concept Indexing. For each of the surgical text words in a case that canonically matched one of the 513 *ICD-9-CM* derived words, a 1 was placed in a corresponding 513 element vector; the vector was set to 0 otherwise. Three vectors were created to each case, one for each text database.

To evaluate the influence of truncation values on retrieval efficiency, we trimmed the solution vectors from the SVD using: 10%, 25%, 33%, 40%, 50%, and 100% of the singular values and vectors. The similarity score (cosine) for each procedure code was computed for each case over all three text datasets and at all seven truncation values of the SVD solution. Thus, for each case, 21 vectors of 259 cosine values were generated from the ICD-9-CM sources, representing the strength of association for that case and each of the 259 procedure codes. These produced 21 matrices of 5,512 cases by 259 concept cosines, one matrix for each pair of text dataset and truncation value combinations.

Inquiries and Correct Retrieval Sets. A set of 15 inquiries related to cardiovascular surgery was drafted, drawing from our experience with over 2,000 clinical research inquiries each year. The queries are listed in Table 1.

Table 1
Evaluation Inquiries

Text	No. Cases	ICD-9-CM Codes
pulmonary valve replacement homograft	1	35.25
repair of Tetralogy of Fallot	12	35.81
repair of atrioventricular canal	17	35.54 35.63 35.73
endarterectomy of carotid artery	221	38.12
repair arteriovenous fistula	17	39.53
repair of secundum atrial septal defect	65	35.71 35.61 35.51 35.52
closure of patent foramen ovale	47	35.71 35.61 35.51 35.52
bilateral aortofemoral or aortoiliofemoral artery bypass graft	101	39.25
axillofemoral bypass graft	18	39.29
thrombectomy radial artery	7	38.03
arteriovenous fistula or shunt for renal dialysis	186	39.27
ligation occlusion superior vena cava	6	38.7
thrombectomy arteriovenous dialysis shunt	90	39.49
ligation occlusion of patent ductus arteriosus	28	38.85
reimplantation of aberrant renal artery	1	39.55

For each query, we used broad ICD-9-CM procedure codes and selected keywords to identify several hundred possible matches from the 5,512 case set. These cases were independently reviewed by two experienced nosologists to identify the correct case set; discrepancies in case selection were arbitrated by a panel of the nosologists and an experienced internist (CGC). The number in parentheses after each query represents our judgement of the correct retrieval set size from the 5,512 cases. Cases were selected with replacement, thus some cases can and do belong to more than one solution set; however, this is mostly attributable to the average of 4.6

coded ICD-9-CM procedures per surgical case.

The words contained in the inquiry phrases were reduced to canonical form, which includes a modest contraction of synonyms at the word level [3]. These canonical words were converted to vector form, analogous to the process for each case. A series of seven cosine value vectors was created for each inquiry at each level of SVD solution truncation. Thus, matrices similar to those created for the surgical text datasets were created for the query set, and provided the basis for a second level of similarity cosine computations.

LSI Retrieval. The matrix of inquiry similarity values (cosines) was computed, vector by vector (inquiry), across the three surgical dataset cosine matrices to yield a second stage vector of cosine similarity values that rank ordered the surgical cases for relevance to each inquiry. This process was repeated seven times for each corresponding matrix set of SVD solution truncations. The algebra of this computation has been previously reported in detail[3].

SMART Comparison. To provide a widely accepted standard of comparison for our retrieval performance, we used the SMART system (Version 10) developed at Cornell by Salton and others [8] on each of our surgical text data sources. Default parameters were chosen in every case. No effort was made to optimize these retrievals, and they should not be interpreted as the best performance capable by the SMART system. Simple text word frequency, and frequency adjusted for discrimination value (rarity $\cong$ inverse term frequency) were used to create two curves.

Surface Matching. To compare LSI performance against less sophisticated techniques, we used the original information matrix in undecomposed form (before SVD) to compute cosine similarity scores. Thus, canonical words were represented as either a 1 or a 0 in the case vector, as opposed to a cosine value deriving from the decomposed and truncated LSI matrix. This allowed us to model the cosine retrieval process on surface matches of words simply reduced to preferred form.

Procedure Code Retrieval. *ICD-9-CM* procedure codes were selected which correspond to each inquiry, these also appear in Table 1. From 1 to 3 codes were required for each query. Because cases retrieved by code were not rankable, Precision-Recall (PR) curves could not be generated, but only the average PR point.

Precision-Recall. We computed standard precision-recall (PR) curves [9] for each combination of retrieval experiments. Each point represents the average over all 15 inquiries, varied as the rank ordered list of retrieved cases was descended. Thus, no threshold values of absolute similarity score were employed *per se*.

Environment. Text processing was performed by PERL [10] scripts and the UNIX shell. Statistical computation was performed using the M++ Math Library of C++ class objects [11]. Graphs were generated by S-Plus [12]. All processing took place on a variety of Sun SPARCservers and workstations under Solaris 1.0.

RESULTS

LSI of the procedure titles produced the PR curve in Figure 1. Only five truncation values among the seven computed are shown for clarity. The truncation fraction did not greatly impact retrieval performance. Figure 2 shows corresponding data for the body of the surgical dictation. The additional text of the full operative report degraded PR performance using this LSI approach. The text data set for both procedure titles and dictation was not materially different from that for the dictation alone (data not shown).

Two curves appear in Figure 3 for the SMART system, the dotted curve invoked a term discriminating function adjusted for term rarity. The dashed curve resulted from the cosine matches based on surface matching. Finally, the large diamond on the right of Figure 3 shows the PR point deriving from the humanly assigned *ICD-9-CM* procedure codes.

DISCUSSION

The superior retrieval performance of a humanly coded standard classification approach relative to these information retrieval techniques is clear. Among the semiautomated approaches, LSI has an apparent disadvantage to the document retrieval systems developed by Salton (SMART), and simplistic surface matching.

The original application of LSI by Deerwester and others was to document retrieval as a function of document to document similarity. No classification system or concept base was invoked. The utility of classifying patient events to conceptual categories, codes, or rubrics is demonstrated by the remarkable retrieval performance of the humanly coded data. Our application of LSI attempted first to classify the text to a code or vector of codes (the cosine vector of partial matches), and then retrieve across these vectors relative to a query. This is therefore a two-step process, and indeed invokes sequential matrix cosine computations.

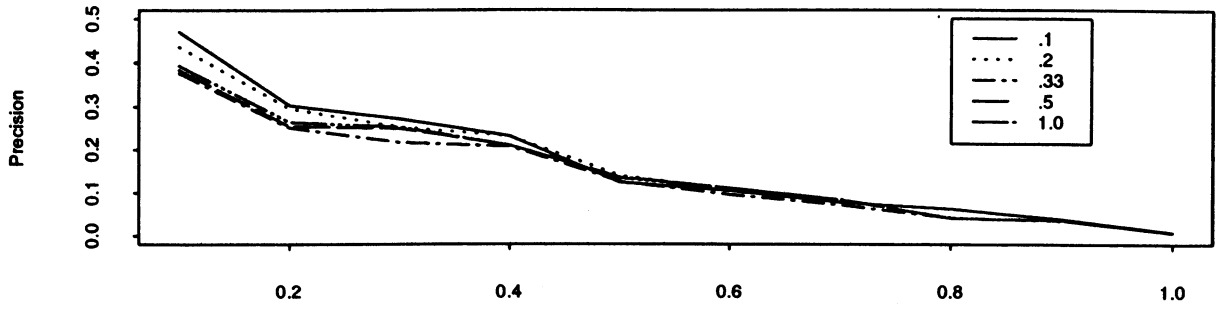

Figure 1: Surgical case retrieval efficiency for procedure title text file. Boxed legend refers to LSI truncation values.

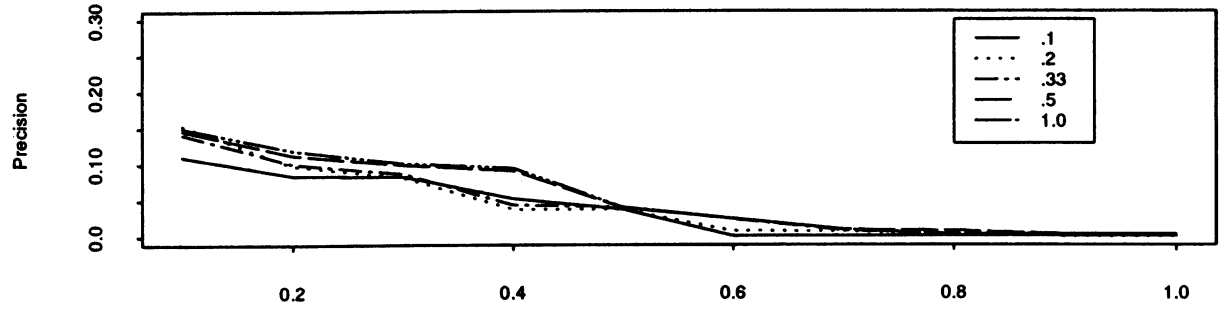

Figure 2: per figure 1 for operative report text file.

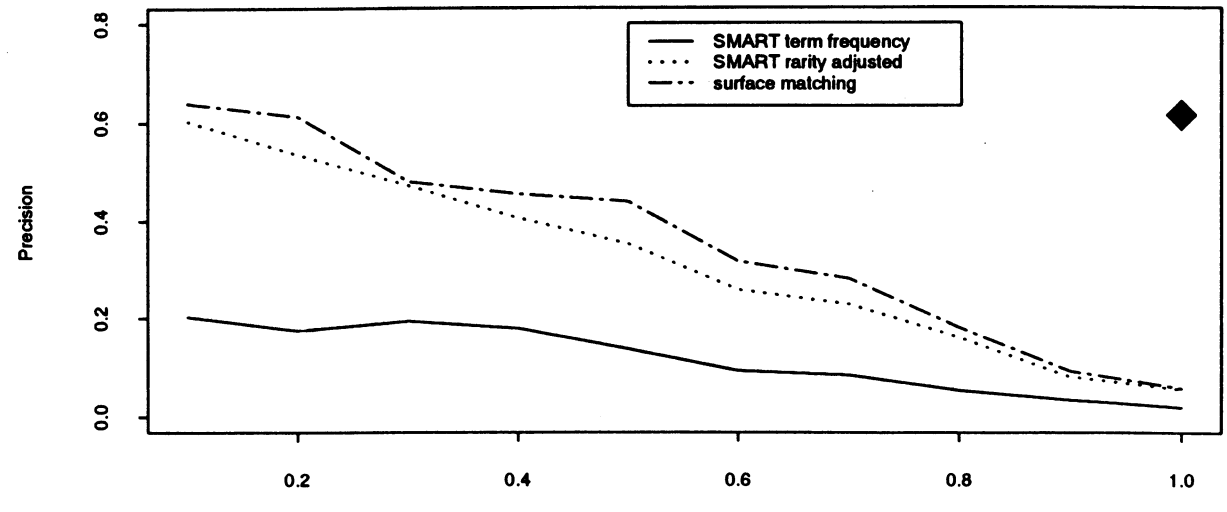

Figure 3: Standard retrieval techniques. Diamond obtained from humanly coded ICD9-CM procedure codes.

The two stage similarity scores (cosines) computed by our LSI application may explain the largest part of retrieval degradation demonstrated by our method relative to the one stage process of the original LSI methods. Further, our first stage computations are dependent upon a valid semantic representation of the concept category (procedure code) in the text associated with the code in the source document. This corresponds to the index entries of the ICD-9-CM in our experiment being required to be a virtually complete synonym space for the procedure codes, a criteria any clinician knows is not true for the ICD-9-CM index.

These experiments raise concern about the utility of concept based LSI which invoke machine readable semantic sources intended for other purposes, in particular the ICD-9-CM. Our earlier experiments with LSI and the UMLS were restricted to short phrase and term matching [3]. The addition of larger amounts of text appears to add more noise than substance, as exhibited by the poorer performance of full operative report dictation versus the more concise procedure titles in the present work.

Two opportunities remain to be explored before abandoning the considerable promise and elegance of concept based LSI. One is to refine and augment the terms and concepts of the contributing semantic sources, as demonstrated by Evans [13]; however, this abolishes the previously stated advantage of avoiding knowledge base construction [3]. The second is to experiment with discrimination values in the population of the information matrices before SVD. Presently, we assign all words to a weight of 1 or 0. Our experience with the SMART system suggests that its performance is improved by invoking a discriminance value.

The features of "partial credit" concept matching, and multidimensional matrix representation (as in our earlier imaginary plane ancestral hierarchy) commend continued evaluation of the LSI approach. Nevertheless, we are also pursuing radical alternatives to LSI, which model word to concept associations based on a linear least squares fit of humanly coded phrases. This approach is introduced elsewhere in these proceedings [14].

Acknowledgements

We thank Donna Ihrke for carefully maintaining our clinical lexicon and for primary nosologic review, Geoffrey Atkin for programming and analytic insight, Maryanne Mathiowetz and the Surgical Index Coding Staff for graciously providing nosologic review, and Karen Elias for manuscript preparation. This work was funded in part by an NIH consortium grant (AR30582).

Bibliography

1 *International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM).* Ann Arbor, MI: Commission on Professional and Hospital Activities, 1986.

2 Côté RA. *Systematized Nomenclature of Medicine (SNOMED).* Skokie, IL: College of American Pathologists, 1982.

3 Chute CG, Yang Y, Evans DA. Latent semantic indexing of medical diagnoses using UMLS Semantic Structures. *Proceedings of the 15th Annual Symposium on Computer Applications in Medical Care,* 1991:185-189.

4 *International Classification of Diseases, 9th Revision, Clinical Modification, Volume 3.* Ann Arbor, MI: Commission on Professional and Hospital Activities, 1991.

5 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.

6 *International Classification of Diseases, 9th Revision, Clinical Modification.* Ann Arbor, MI: Commission on Professional and Hospital Activities, 1980.

7 Golub CH, Van Loan CF. *Matrix Computations, Second Edition.* Baltimore, MD: The Johns Hopkins University Press, 1989.

8 Buckley C. Implementation of the SMART Information Retrieval System. *Technical Report No. 85-686.* Ithaca, NY: Cornell University, 1985.

9 Salton G. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Reading, MA: Addison-Wesley Publishing Co., 1989, p. 248.

10 Wall L, Schwartz RL. *Programming Perl.* Sebastopol, CA: O'Reilly & Associates, Inc., 1990.

11 *M++ Class Library, Release 3, Users Guide.* Bellevue, WA: Dyad Software Corporation, 1991.

12 *S-Plus Reference Manual.* Seattle, WA: Statistical Sciences Inc., 1991.

13 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.

14 Yang Y, Chute CG. An Application of a Least Squares Fit Mapping to Clinical Classification. *Proc 16th Ann Symp Comp Applic Med Care 1992,* in press.