A Comparison of Intelligent Mapper and Document Similarity Scores for Mapping Local Radiology Terms to LOINC

Daniel J. Vreeman, PT, DPT and Clement J. McDonald, MD
Regenstrief Institute, Inc. and Indiana University, Indianapolis, IN

ABSTRACT

We developed a program for mapping local radiology system terms to LOINC that returns a ranked list of candidate LOINC codes based on document similarity scores. We compared the performance of this program with the Intelligent Mapper (IM) program in mapping diagnostic radiology terms to LOINC. The cosine similarity score ranked the correct LOINC code first in 34% of the terms in our development set and 39% of the terms from our test set, compared with IM's ranking of the correct LOINC code first in 83% of the terms in our development set and 92% of the terms in our test set. This study demonstrates the challenges in using document similarity scores for mapping to LOINC. Because vocabulary mapping is a resource-intensive step in integrating data from disparate systems, we need continued refinement of automated tools to help reduce the effort required.

INTRODUCTION

A fundamental challenge to the goal of interoperable health information exchange is the myriad, idiosyncratic conventions for identifying similar concepts in separate electronic systems. Logical Observation Identifiers Names and Codes (LOINC®) is a universal code system for identifying laboratory and other clinical observations.[1] Mapping local observation terms to LOINC provides a bridge across the many islands of data that reside in disparate systems, but the work of mapping to standardized vocabularies is a barrier to their adoption because of the time and effort it requires.

The Indiana Network for Patient Care[2] is an early example of an operational local health information infrastructure (LHII); it caries hundreds of millions of entries from five different health systems. In this collaborative, we have coalesced many of the various sources that produce and store data in our community, including hospitals, laboratories, and radiology centers. We accomplish the task of integrating data from all of the contributing source systems by mapping the idiosyncratic local terms to a common dictionary based on LOINC. The process of mapping these local terms requires substantial manual effort, domain expertise, and may also be subject to the inconsistencies inherent in human review. We attempt to solve these problems by developing automated tools that improve the efficiency and consistency of mapping.

We previously developed and reported[3] on a tool called Intelligent Mapper (IM), which automatically identifies a ranked list of candidate LOINC codes for each local term in a set. IM generates its ranked list by determining word matches between the local test names and formal LOINC names, with an option to narrow the search space of candidate LOINC terms based on the CPT® codes* that are usually contained in radiology system master files.

IM's success in mapping depends upon ongoing, labor-intensive maintenance of domain-specific synonymy in the LOINC database and a LOINC to CPT linkage table. Because radiology systems generate a large corpus of text documents, we hypothesized that we could identify radiology test names by the distribution of words in the reports, and thus use this as an alternative to test name matching for mapping to LOINC. An approach that leverages the inherent content of the documents would avoid the effort of maintaining a rich synonymy set for words in term names from each domain of interest. A number of information retrieval techniques exist for searching text documents. One commonly employed method is the vector space model (VSM).[4,5] In this model, documents are represented as a vector of terms. Calculating the cosine of the angle between two document's vectors gives a measure of how similar to each other the documents are. While this approach is widely used for retrieving documents, it has not been studied for mapping narrative test terms. We developed a VSM-based program for mapping local diagnostic radiology terms to LOINC, and here report a preliminary, comparative analysis with IM.

METHODS

Vector Space Model

In the vector space model (VSM)[4,5] of information retrieval, documents and queries are stored as sets, or vectors, of terms. The index terms can be words, word stems, concepts, or phrases[6] that are typically

* CPT is a registered trademark of the American Medical Association

obtained from the texts of interest. The degree of similarity between documents can be assessed by calculating the cosine of the angle between two document vectors. For two documents *D* and *E*, we let *Ds* and *Es* be the set of terms occurring in *D* and *E*. We let *T* be the union of *Ds* and *Es*, with *ti* being the *i*-th element of *T*. Thus, we can represent the term vectors of *D* and *E* as:

$$Dv = (nD(t1), nD(t2) \ldots, nD(tN)$$
$$Ev = (nE(t1), nE(t2) \ldots, nE(tN)$$

where nD(ti) is the number of occurrences of term *ti* in the document *D*, with the same being true for nE(ti) in *E*. The cosine similarly of these documents can be calculated with:

$$\cos(Dv, Ev) = \frac{Dv \bullet Ev}{Norm(Dv) * Norm(Ev)}$$

where • is the dot product and Norm is the Euclidean norm (square root of the sum of squares). Because not all terms are equally able to distinguish among documents in the corpus, vector terms can also be assigned a weight to reflect these differences. Term weights are often calculated using the term frequency, inverse document frequency (TF-IDF):[4,5]

$$\text{tf-idf} = tf \bullet \log_2\left(\frac{|D|}{|(d_j \supset t_j)|}\right)$$

where $|D|$ is the total number of documents in the corpus and $|(d_j \supset t_j)|$ is number of documents where the term t_j appears. The TF component measures the importance of the term within the document; the IDF component offsets this by how common the word is in the entire corpus.

We developed a VSM-based program in Perl that returns a ranked list of candidate LOINC codes for each local radiology term in a set, based on the similarity scores between the local terms' documents and the LOINC codes' documents. We implemented the VSM using a freely available Perl module, Text-Document-1.07 (http://www.cpan.org/).

Normalizing Terms

The VSM method is often built on the collection of terms in the corpus of interest. The words used in radiology narratives display considerable morphological variation that could hinder such an approach. To combat this variation, we used a normalization program, LuiNorm, developed by the National Library of Medicine as part of the Lexical Tools package[7] and implemented in Java. LuiNorm returns a canonicalized form of any input word by abstracting away from case, inflection, stop words, genitive markers, punctuation, diacritics, ligature, and word order. For example, given either of the input words "regarding" or "regards", LuiNorm returns the canonicalized form "regard".

LOINC, RELMA, and the Intelligent Mapper

Our previous work[3] of evaluating the Intelligent Mapper program was done in conjunction with an expansion of LOINC content for radiology report names. We have continued this expansion effort, having added over 800 new terms since that report. In our previous analysis, IM's best performance was achieved when it used a feature to restrict the search for candidate LOINC codes among those that shared a CPT code with the local term. In that project, we created a LOINC to CPT mapping table for use in IM, and have continued to update this table as new content is added to LOINC. For this study, we mapped local terms to our in-house version of the LOINC database, version 2.16+, which contains 340 more radiology terms than the last public release.

The Regenstrief Institute, Inc. has developed a program called the Regenstrief LOINC Mapping Assistant (RELMA) program that is distributed with each LOINC release and made freely available (http://loinc.org). RELMA contains tools for both browsing the LOINC database and mapping local test codes to LOINC. RELMA's basic interface facilitates term-by-term mapping of local terms to LOINC by providing search capabilities that return candidate LOINC codes based on keywords given by the user. RELMA also contains the Intelligent Mapper program. For this study, we used last public release, version 3.16, of the RELMA program.

Data Sources

We used the HL7 message streams from two institutions and the repository of our LHII[2] to assess the performance of document similarity scores and IM. This study was approved by our local institutional review board. For our *development set*, we extracted one month of diagnostic radiology reports from the HL7 message stream of the urban not-for-profit hospital that served as our development institution in our prior analysis of IM. From the one month extract, we chose 50 diagnostic radiology tests that had more than one report sent during that month. We used this sample of terms and reports for developing our VSM-based program. Additionally, this institution's tests were used in the development of IM and LOINC terms in our prior work.

For our *test set*, we extracted two months of diagnostic radiology reports and terms from HL7 messages of a local, large physician group practice. We used all of the terms and reports in this sample as

our test set for evaluating the VSM-based program and IM, because we had not used these terms in developing either of these programs.

For both our development and test sets, the extract provided the local radiology test names and codes and a corpus of narrative reports. Our test set contained 2,375 messages for 200 radiology test codes. As is often the case in radiology system terms, this set contained different local codes for the same test done at different facilities within that health system. We pre-processed the local terms with a Perl script to squeeze the set down to only contain records with unique test names. Because this analysis and the LOINC expansion effort have focused on diagnostic radiology tests, we excluded terms that represented interventional radiology or nuclear medicine tests. These consolidations left us with 2,125 diagnostic radiology reports representing 150 terms. We then excluded any reports whose entire narrative content consisted of "Please refer to examination dated <date>" with no other clinical content. Our final test set thus contained 1,952 reports for 143 terms.

We used the mappings to LOINC from our LHII[2] to extract sample reports for as many diagnostic radiology LOINC codes as possible. We limited the number of extracted reports to a maximum of 250 for each of the 716 LOINC codes in our mappings. This extract yielded a total of 116,536 reports (mean number of reports per LOINC code = 163, mode number of reports per LOINC code = 250).

Gold Standard Mapping

A domain expert (DJV) manually established a gold standard mapping to LOINC for both term sets against which to compare the VSM-based program and IM's results. The mapping process followed the recommended procedures outline in the RELMA User's Manual[8] and previously described.[3]

Our mapping rules stated that the gold standard mapping would be an exact correspondence from the local term to LOINC. If no LOINC match was identified, we counted that local term as unmapped. As a component of the standard mapping procedure, we identified the words in our local term names unknown to LOINC, and translated them into known LOINC words where possible. For example, we translated the word "lmboscrl" into "lumbosacral", the word that LOINC knows. We provided translations for 61 words in our development set and 37 words in our test set. In the final step, the domain expert used RELMA to search for LOINC code matches on a term-by-term basis. Reliability for manual mapping was not established.

VSM Processing for Mapping

For each set of extracted radiology reports (development set, test set, and LOINC code set), we pre-processed the narrative body of the report with a Perl script to lowercase all words, remove non-word characters, remove excess whitespace, and to strip out any institution-specific header or footer information. We then normalized the narrative text using the LuiNorm[7] program, setting its parameters to remove tokens less than 2 characters long.

We removed the word "with" from LuiNorm's list of stopwords because it is commonly used in radiology reports to indicate the presence of contrast (e.g. "mri of liver was performed *with* intravenous contrast"). After reviewing the list of words in our development set, we added 20 more words to the default list of 9 stopwords used by LuiNorm. We also removed the institution-specific section header labels (e.g. "admitting diagnosis", "impression"). After removing these stopwords, we created a "document" for each of the terms in our three sets as the aggregation of all the radiology reports extracted for that term.

We ran our VSM-based program on both the development set and test set, comparing the similarity between the documents of the local terms and the documents of the LOINC codes. The program calculates a cosine similarity score and a weighted cosine similarity score, based on the TF-IDF weightings. The outputs of the program are two ranked lists of candidate LOINC terms for each local term, one based on the cosine similarity score and the other based on the weighted cosine similarity score.

The VSM-based program and LuiNorm were run on a computer with dual Athlon MP 1900 processors and 4.0 GB of RAM, using the RedHat Enterprise Linux 4 operating system.

Intelligent Mapper Processing for Mapping

We also processed the terms from both our development and test sets with IM to identify candidate LOINC terms. We used IM's CPT-based restriction for all 5 digits of the CPT code, because we had previously found this to be the most accurate.[3] We used the vocabulary translations for unknown words that were identified in the gold standard mapping, and selected the user option to limit the search to only LOINC codes in the "radiology studies" class.

The gold standard mapping and Intelligent Mapper analyses ran on a 1600 MHz computer with 1.0 GB of RAM, using the Windows XP operating system.

Measures

For both the VSM-based program and IM, we calculated the program's ability to include the correct LOINC code in its top ranking and recorded computational cost. We limited the list of candidate LOINC codes to the top five, because our previous work with IM indicated that few additional matches were found in between ranks five and ten. We evaluated both programs' accuracy for identifying correct matches and describe these findings in the context of the clinical message flows.

RESULTS

The gold standard mapping identified a true LOINC match for all 50 terms in our development set and for 130 of the 143 terms in our test set. Of the true LOINC matches for terms in our development set, we could extract radiology reports from our LHII for 41 of the 50. Similarly, we could extract radiology reports from our LHII for 104 of 130 the true LOINC code matches in our test set. Table 1 gives the performance of IM and the VSM program for identifying the LOINC code matches in these terms where it was possible to calculate a similarity score for the true LOINC match.

Overall, IM more accurately identified correct LOINC matches than did either document similarity score for both term sets. In both our test set and our development set, there was no significant difference in success of matching between the cosine similarity score and the weighted cosine similarity score for ranking the correct LOINC code first, in the top three, or in the top five ($\chi^2 P>0.05$). In our test set, the difference in success of matching between IM and either similarity score was significant for ranking the correct LOINC code first, in the top three, and in the top five ($\chi^2 P<0.0001$). We did not calculate the χ^2 value for matching success between IM and the document similarity scores in our development set, due to the small sample size. Because these results were calculated only for terms where the gold standard LOINC was present in our data extract, they represent the document similarity score's recall (correct matches made / correct matches possible).

The computational cost of the VSM-based program was higher than that of IM. The processing time to return a ranked list of top five candidate LOINC codes for our development set was seven minutes for IM, and three hours 10 minutes for the VSM-based program. The processing time to return a ranked list of top five candidate LOINC codes for our test set was one hour 20 minutes for IM, and six hours 30 minutes for the VSM-based program.

Table 1. Performance of document similarity scores and Intelligent Mapper for identifying correct LOINC matches.

	Correct LOINC Codes Returned		
	Top Ranked % (n)	**Rank in Top 3** % (n)	**Rank in Top 5** % (n)
Development Set (n=41)			
Document Similarity Score			
Cosine	34 (14)	68 (28)	85 (35)
Weighted Cosine	39 (16)	71 (29)	83 (34)
Intelligent Mapper	83 (34)	90 (37)	95 (39)
Test Set (n=104)			
Document Similarity Score			
Cosine	39 (41)	67 (70)	75 (78)
Weighted Cosine	38 (39)	62 (64)	71 (74)
Intelligent Mapper	92 (96)	97 (101)	97 (101)

The two month extract of HL7 messages for our test set contained 1,952 diagnostic radiology reports. If the LOINC code that the best performing document similarity score (cosine) ranked first was assigned to the radiology report codes from this physician group, 25% of the reports in the message extract would be correctly mapped. If LOINC codes were assigned to reports in this way based on IM's top ranking, 94% of the reports would be mapped correctly. The cosine similarity score ranked the correct LOINC code in the top three for 62% of the reports, and ranked the correct LOINC code in the top five for 64% of the reports. IM ranked the correct LOINC code in the top three and in the top five for 95% of the reports.

DISCUSSION

The VSM-based similarity scores did not identify the correct LOINC code matches for our local radiology terms as accurately as IM did. There are several factors that may be contributing to the performance of the VSM-based similarity scores in our study. Our determination of mapping success was based on an exact match between the local term and the LOINC code, because this is the level of precision we typically expect in mapping test codes for clinical data exchange. Many evaluations of VSM-based systems assess the "relevance" of the returned documents to the input query. Compared to a typical information retrieval evaluation, our "exact term match" operational definition of relevance was stringent, but pragmatic.

We had expected that the weighted cosine similarity metric would outperform the cosine similarity metric in returning candidate LOINC codes, but instead we found no difference. The TF-IDF method gives a high weight to terms occurring frequently in the document but rarely in the rest of the corpus. In reviewing the ranked list of candidate LOINC codes, we noted that failure to disambiguate the appropriate

use of a contrast agent (e.g. "with contrast", "without contrast", or "with and without contrast") was a frequent reason for returning an incorrect match. Commonly occurring words in the corpus such as "with" and "without" will not be highly weighted by TF-IDF. In future work, we intend to investigate the utility of other term weighting schemes.

We canonicalized the narrative content of the extracted reports to reduce the inherent variability. We used the freely available LuiNorm tool to achieve this abstraction; however, it is possible that other stemming algorithms[9] or methods for term[10] or concept[11] identification may perform better for this purpose. For example, autocoding to a standard nomenclature[12] could be used to identify the index terms used in the VSM. Although stem-based VSM is widely used, Mao and Chu[6] have reported improved retrieval accuracy using phrase-based VSM compared to word stem-based VSM.

The radiology reports in our test set were generated largely in the context of outpatient care, whereas the reports extracted for the LOINC codes from our LHII likely contained a higher proportion of studies performed in the inpatient setting. Differences in the report content for the same clinical test performed in these two care settings could have reduced the accuracy of the VSM-based similarity scores. Despite these differences, we wanted to evaluate the VSM approach in the real-world context of mapping for an LHII, where the proportion of tests done in these settings may vary between participating institutions.

We could extract sample reports for 80% of the gold standard LOINC codes for our test set terms, but the fact that we could not extract reports for all of them highlights a limitation of the VSM approach in requiring a corpus of documents for both source and target mapping terms. Our test set of terms was built from two months of HL7 messages; however, more than two thirds of these tests had fewer than 10 reports in that time period. The small number of reports for these tests also could have reduced the similarity scores' ability to identify correct matches.

We expected the higher computational intensity that we observed for the VSM program compared to IM. The VSM program operates on the aggregate narrative content of all reports for each term, whereas IM operates on only the test name words. Because the VSM program runs unaided, the increased computational time is likely affordable.

Intelligent Mapper identified correct LOINC matches for the terms in our test set with accuracy similar to our previous analysis.[3] By replicating these findings in an independent sample from another institution, the current study lends support to the generalizability of IM's high accuracy in the domain of diagnostic radiology. We were able to identify a gold standard LOINC mapping for 91% of the terms in our test set, which evidences the relative completeness of LOINC's coverage for diagnostic radiology tests.

CONCLUSION

This study demonstrates the challenges in using information retrieval methods for mapping local radiology system terms to LOINC. The VSM similarity scores did not identify the correct LOINC codes as accurately or as efficiently as IM did. Because vocabulary mapping is a resource-intensive step in integrating data from disparate systems, we need continued refinement of automated tools to help reduce the effort required.

References

1. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin Chem*. 2003;49(4):624-633.
2. McDonald CJ, Overhage JM, Barnes M, et al. The Indiana network for patient care: a working local health information infrastructure. *Health Affairs*. 1005;24(5):1214-1220.
3. Vreeman DJ, McDonald CJ. Automated mapping of local radiology terms to LOINC. *Proc AMIA Symp* 2005:769-773.
4. Salton G, McGill MJ. Introduction to modern information retrieval. McGraw-Hill, 1983.
5. Salton G, Buckley C. Global text matching for information retrieval. *Science*. 1991;253(5023):1012-1015.
6. Mao W, Chu WW. Free-text medical document retrieval via phrase-based vector space model. *Proc AMIA Symp 2002*:489-493.
7. National Library of Medicine. Lexical Tools. Available at: http://umlslex.nlm.nih.gov. Accessed March 2006.
8. LOINC Committee. RELMA® version 3.16 user's manual. Indianapolis, IN:Regenstrief Institute, 2005. Available at: http://loinc.org Accessed March 2006.
9. Lovins JB. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*. 1968;11(1):22-31.
10. Krauthammer M, Nenadic G. Term identification in the biomedical literature. *J Biomed Inform*. 2004;37(6):512-26.
11. Huang Y, Lowe HJ, Klien D, Cucina RJ. Improved identification of noun phrases in clinical radiology reports using a high performance statistical natural language parser augmented with the UMLS Specialist Lexicon. *JAMIA*. 2005;12(3):275-285.
12. Berman, JJ. Doublet method for very fast autocoding. *BMC Med Inform Decis Mak*. 2004;15(4):16.

Acknowledgements

The authors thank Kathy Mercer for LOINC development and Tammy Dugan for data extraction. This work was performed at the Regenstrief Institute, Inc and was supported in part by the National Library of Medicine (N01-LM-3-3501).