2011-06-30

Performance Analysis and Assessment of a TF-IDF Based Archetype-SNOMED-CT Binding Algorithm

Sheng Yu
Dublin Institute of Technology, sheng.yu@dit.ie

Jesus Bisbal
Universitat Pompeu Fabra, Barcelona

Damon Berry
Dublin Institute of Technology

Recommended Citation

Yu, S., Bisbal, J. & Berry, D. (2011). Performance Analysis and Assessment of a TF-IDF Based Archetype-SNOMED-CT Binding Algorithm. *24th International Symposium on Computer-Based Medical Systems (CBMS)*, Bristol, UK. doi:10.1109/CBMS.2011.5999029

Performance analysis and assessment of a *tf-idf* based Archetype-SNOMED-CT binding algorithm

Sheng Yu
Dublin Institute of Technology
TeaPOT Research Group
School Elect. Eng. Systems
Kevin St., Dublin 8, Ireland

Damon Berry
Dublin Institute of Technology
TeaPOT Research Group
School Elect. Eng. Systems
Kevin St., Dublin 8, Ireland

Jesus Bisbal
Universitat Pompeu Fabra
Department of ICT
Barcelona, Spain

Abstract

Term bindings in archetypes are at a boundary between health information models and health terminology for dual model-based electronic health-care record (EHR) systems. The development of archetypes and the population of archetypes with bound terms is in its infancy. Terminological binding is currently performed "manually" by the teams who create archetypes. This process could be made more efficient, if it was supported by automatic tools. This paper presents a method for evaluating the performance of automatic code search approaches.

In order to assess the quality of the automatic search, the authors extracted all the unique bound codes from 1133 archetypes from an archetype repository. These "manually bound" SNOMED-CT codes were compared against the codes suggested by the authors' automatic search and used for assessing the algorithm's performance in terms of accuracy and category matching. The result of this study shows a sensitivity analysis of a set of parameters relevant to the matching process.

1. Introduction

The harmonisation of clinical data models and terminology models is driven by advocates of both modelling methodologies. One major challenge associated with this work is the need to annotate the clinical information with an appropriate concept from a terminology to bring portability and interoperability. Recent research interest on this topic has focused on using an automatic means to annotate clinical information in an electronic health record with concepts from external terminology. The difference between annotating free text in clinical notes and tagging clinical concepts in modern EHRs is that data within modern EHRs are organised by a structurally constrained information model.

One information modelling approach that has been gaining momentum employs so-called *archetypes* [4] which express the views of clinical experts and contain both structural and semantic constraints. SNOMED-CT is a large medical terminology system. Members of the openEHR organisation are working on providing links to SNOMED-CT terms within archetypes and also to harmonise SNOMED and openEHR representations in order to deliver enhanced semantic interoperability in e-health. A number of studies [8] [7] have already been conducted under the assumption that automatic annotation with SNOMED-CT produces sound and reliable outcomes. However there are few publications in the literature that report the effectiveness of automatic SNOMED-CT binding methods in terms of accuracy and reliability. Binding methods of this type should also adapt to the constantly developing SNOMED-CT to suit particular clinical scenarios.

Hence it is necessary to quantitatively evaluate the performance of these annotation algorithms. This paper builds upon a previously published *tf-idf* based

SNOMED-CT binding approach [11] and performs a sensitivity analysis to evaluate the effectiveness of the automatic binding process. The assessment can be extended to facilitate and promote the continuous improvement of binding algorithms in future work.

2. Background

An Archetype model could be considered as a type of clinical meta-data model. Archetypes are clinician-created meta-information models that describe constraints for data stored in an EHR. The archetype-based approach is a promising EHR paradigm because with archetypes, medical knowledge can be separated from technical implementation. Due to the nature of community-based development, archetypes are created according to agreed medical definitions by archetype modellers. Repositories with large numbers of archetypes are being created and the difficulty associated with managing archetypes is increasing. In the future, the large amounts of clinical content described by archetypes will need to be properly categorised and users should have easy access to them [5].

On the other hand, SNOMED-CT is a multi-purpose clinical vocabulary that consists of hundreds of thousands carefully modelled medical concepts. Its substantial coverage of clinical content and phenomena makes it a highly-rated external terminology for disambiguating and clarifying clinical statements. The hierarchical structure of SNOMED-CT means that it can be used as a classification system for various purposes.

The association of archetype terms with external terminology improves the semantic interoperability during communication between different health organisations. While archetypes contain their own definitions of concepts, they can refer to formal terminology system such as SNOMED-CT thus to reduce the ambiguity in clinical data. Binding to commonly understood SNOMED-CT terms facilitates common understanding between diverse EHR systems and so promotes semantic interoperability. Other benefits of annotating archetypes with external terminology have been discussed elsewhere [10] [2]. However, due to the large quantity of archetypes being created, purely manual annotation is not appropriate to obtain relevant and high quality bindings. Some form of automated support is necessary. Sound algorithms need to be created to find appropriate SNOMED-CT concepts.

3. Related work

Lezcano [7] et. al., used a UMLS thesaurus utility, comprising of a normalised string search to associate archetype nodes with SNOMED-CT concepts. The normalised search compares input strings with records of a large built-in string index. The process of building indices of this type involves normalising terms in a UMLS thesaurus so that strings can be compared in a normalised way i.e ignore word sequence, tense etc. Similar to a database keyword look-up, the output of this type of binding process is largely dependent on its index. Our experience of this tool is that the input string has to be quite similar to the record for them to match. The system often fails to produce any match if the input query contains more words, given that some word might not appear in the index.

Qamar [9] et al. adopted a multiple searching and filtering approach to associate archetype nodes with SNOMED-CT codes. Multiple natural language processing tools such as GATE wordsense disambiguation were used to aid the binding process. The filtering rules were chosen based on judgement and then assessed by a group of experts. However the assessment is not easily scalable and it would be difficult to apply this algorithm to other related research. By comparison, because the assessment relies on a resource created by experts rather than the experts themselves, the contribution presented here provides more quantitative and re-applicable analysis of the binding method used to annotate archetypes.

4. Algorithm evaluation

This paper aims to provide an initial demonstration of an evaluation technique as an example to support future studies which may utilise binding-assistance algorithms of this type. The rating approach is applicable for more complex algorithms. One reason for evaluating the comparatively simple algorithm shown here is that the resources required for the algorithm are generally available for ubiquitous demonstration. The "gold standard" that is used to assess the "effectiveness" of the automatic binding suggestion, the bound codes in archetypes, is publicly available for this method. Another reason, is that the method is convenient (using a single tool Lucene) and transparent.

The previous work utilised a free text indexing tool called Lucene[6] featuring the term frequency-inverse document frequency (*tf-idf*) weighting scheme to enable the ranking of search results. The *tf* factor is the frequency of a term inside a document and the *idf* factor is the inverse of the frequency of a term among the documents in the collection[1]. It was adapted in the reported work to provide an automatic means to search by archetype terms and suggest their best matches in SNOMED-CT. The approach adopts the following techniques, 1.All SNOMED-CT terms are indexed to enable searching by free text. 2.Archetype terms are extracted and queried against the index. 3.Results are then generated to suggest a list of SNOMED-CT terms.

This automated binding mechanism could be seen as an information retrieval (IR) system. In authors' view, in order to use the generated binding suggestions, an analysis of the underlying algorithm needs to be conducted to quantitatively demonstrate its accuracy. Precision and Recall [1] are used extensively to evaluate IR algorithms. *Recall* is the fraction of relevant items that have been retrieved over all relevant items, which can be attached to the question. This reflects how many relevant items the algorithm managed to collect. *Precision* is an indication of the fraction of relevant items in a result set. It reflects the accuracy of the algorithm.

5. Method of analysis

The evaluation of Information Retrieval methods typically feature a collection of documents and a set of queries, which are used to seek the "relevant" documents. Also, a set of known answers associated with these queries indicate whether the document is relevant to the query or not. All testing documents, queries and answers are peer-reviewed by experts, in order to test new IR methods. Examples of such standard document collections include the Text Retrieval Conference (TREC) collection for general free text document retrieval and the Cystic Fibrosis collection for medical text retrieval [1]. In this paper, analogously to information retrieval, a "gold standard" is established by extracting existing manually-selected codes from the set of archetypes. These existing manually-bound SNOMED-CT codes are used as a reference to decide whether a search result has been successful or not.

Because the returned answers will be ranked, this study applied thresholds to trim the number of answers, which were set at a range of different values to establish the impact on the results. Observation of performance related factors was made according to different thresholds. The details of these two different thresholds are described below.

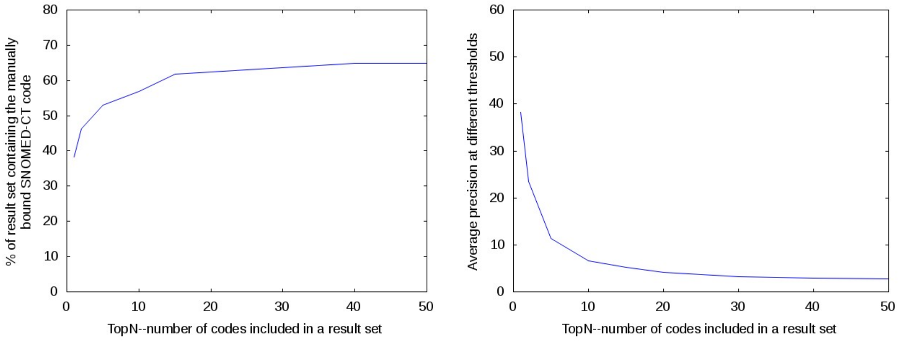
- **TopN** a threshold to set the algorithm to gather the maximum number of collect-able answer(code)s
- **MAJ** a threshold to define at least how many in a result set which are of the same category can be considered the "majority"

6. Result

By setting the threshold to an intended range at suitable data points, the following diagrams in Figure 1 were generated to show how the algorithm performed by measuring the average recall and precision of all the searches[11].
One may argue that under such conditions it is always a 1:1 mapping (1 archetype node is mapped to 1 SNOMED-CT code, thus the number of relevant codes is 1) so the recall can be either 0 or 1. The precision would be 0 or 1 divide by the number of codes in the result set. The diagrams are not in the conventional "precision versus recall" style because both are changing when different thresholds are applied. However the value of calculating these figures can be seen in the future when proper data sets are complete in which case an archetype node may be associated with many SNOMED-CT candidates 1:* (based on experts' judgement). Despite the small size of existing bound codes, this study can still reveal the capabilities of whether it retrieves the only relevant code or not.
Figure 1(a) shows the average recall of the algorithm i.e the percentage of automated results that contain matches for manually bound codes. The trend shows that when more codes are retrieved, it is more likely that the result set will include the code suggested by the "gold standard". Although the average recall remained almost the same after the result set contained 20 and more codes. However more codes in a result set also indicate more "noise" i.e lower precision.

Figure 1(b) shows the average precision of total searches using the algorithm with different thresholds at recall 1.0. The highest average precision value was achieved at threshold value 1 which indicates the result set consisted of only one code, i.e the top answer. The reason for its rapid decline in the range of 1 to 10 is because although more codes were returned in the result set, the chance of retrieving the relevant code did not increase greatly. Only a few more searches retrieved the relevant code when the size of the result set is bigger than 1. Therefore it is suggested that if this version of the algorithm is to be used to automate the binding process, 38.27% of its top answers can be considered as reliable references. Despite the low accuracy of this algorithm, better algorithms can be designed based on a more mature data set.

Another type of comparison was conducted, which took the majority of a result set and checked whether



(a) Average recall

(b) Average precision

Figure 1: *recall* and *precision* versus number of codes in result sets

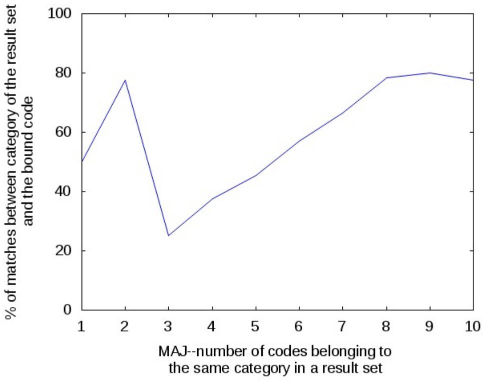


Figure 2: Comparing category of the majority and the bound code

their SNOMED-CT category matched the category of the bound code. The vertical axis in Figure 2 refers to the number of any majority of returned codes in one result set having the same category. The horizontal axis represents the number of codes belonging to the same category in a result set. For instance, where the result set has 9 codes having the same category, the chance would be 77.7% for it to match the category of the bound code. The first 1,2,3 implies few codes shared a category. The category is taken randomly from any group of codes to match the bound code's. Where the majority codes dominated the result set, it is more likely to match the bound code's category in SNOMED-CT. The first two figures illustrate that generally as more answers are included in a result set, it is more likely to include the manually selected code. The percentage of matches is an indicator that when results of the search are used, only a fraction of them are accurate. The third figure, however, shows that besides examining exact matches, other interesting features of the algorithm can be discovered. It reveals that there is a stable growth of category matching between the bound code and the majority of the result sets. It therefore shows the potential of using this algorithm to classify unbound archetype nodes.

7 Discussion

Many interpretations could be used to explain the behaviour of the algorithm as shown in the diagrams. However to a great extent the underlying tf-idf scheme is playing an important role in the retrieval and ranking of codes. Although a review of the tf-idf formula and the vector-space model is not the focus of the paper, insight of how results are ranked can benefit future improvement. Equation 1 is the simplified formula for calculating the scores of each returned result.

$$score = \sum_{t} (tf * idf * FieldFactor). \qquad (1)$$

The field factor is a score to give higher weight to shorter terms. Equation 2 and equation 3 show how tf, idf factors are calculated respectively for each query.

$$tf = termFrequency^{\frac{1}{2}}. \qquad (2)$$

$$idf = 1 + \log(\frac{numDocs}{docFrequency + 1}). \qquad (3)$$

e.g A search for "blood" will retrieve "Blood" as the top answer in a result set containing the following score in total 1068278 SNOMED-CT terms: 5.5406284. By checking the factors from the result set, the following scores were obtained : tf=1.0; idf=5.5406284; FieldFactor=1.0 (docFreq=11394, numDocs=1068278) In this scenario, the number of words in a term is quite small compared to document-like articles. Due to this fact, term frequency tf tends to be 1.0 i.e no repeated word in a single term. From the resulting score the factor with the highest impact seems to be the inverse document frequency. This indicates that if a word appears to be "common" which frequently occurs in SNOMED-CT terms, then it will contribute less to the final score. However other factors such as the field factor, matter also. In the example given above, the top answer consists of only one word "blood" which makes the field factor much larger comparing to other answers with more words. As a result of this basic conclusion, it is suggested that since the idf factor is significant in ranking and retrieving terms, a study of total terms in SNOMED-CT and its distribution can help to improve the results.

8. Future work

In the improved version of the binding algorithm, various natural language processing techniques will be added [3] such as synonym recognition and word disambiguation among archetype terms. The structure of archetypes and reference model information could also be made to influence future algorithms, by incorporating information about the reference model class and archetype path associated with an archetype term to enhance the binding process. There are other SNOMED-CT related changes that could be made, for example, if an archetype is related to "Procedure", flags should be sent to indicate the results should give higher relevance to "Procedure" hierarchy and its related concepts. An important feature of SNOMED-CT, *post-coordination*, needs to be considered when no result is suggested by the algorithm but also in some cases where results are returned. However this use of SNOMED-CT is rather advanced and requires more domain knowledge so future development of automatic post-coordinating SNOMED-CT codes should proceed with caution.

Future work also involves investigating common words in SNOMED-CT world and how their synonyms are distributed. Because of the significance of idf, the subsetting of SNOMED-CT terms could contribute to specialised searching. This could also be combined with other modifications of the algorithm, such as boosting certain factors in the indexing and searching phase, in order to retrieve more desired SNOMED-CT codes.

9. Conclusion

This work has demonstrated that the authors binding algorithm exhibits an acceptable level of performance. The evaluation approach reported here can also be reapplied to analyse future potential automatic binding algorithms and provide a means to assess them. It is hoped and expected that the performance of different algorithms can be compared in this way. The authors believe that this study can be the foundation of exploiting IR and other medical text processing techniques to benefit the archetype-SNOMED-CT binding process.

References

[1] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[2] J. Bisbal and D. Berry. Archtype alignment:a two-level driven semantic matching approach to interoperability in the clinical domain. In *Proceedings of the International Conference on Health Informatics, HEALTHINF 2009*, pages 216–221. INSTICC Press, 2009.

[3] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5):392–402, 2004.

[4] S. Garde, E. Hovenga, J. Buck, and P. Knaup. Expressing clinical data sets with openEHR archetypes: A solid basis for ubiquitous computing. *International journal of medical informatics*, 76:S334–S341, 2007.

[5] S. Garde, E. Hovenga, J. Granz, S. Foozonkhah, and S. Heard. Towards a repository for managing archetypes for electronic health records. In *HIC 2006 and HINZ 2006: Proceedings*, page 61. Health Informatics Society of Australia, 2006.

[6] O. Gospodnetic and E. Hatcher. *Lucene in Action (In Action series).* Manning Publications, Dec. 2004.

[7] L. Lezcano, S. Sánchez-Alonso, and M.-A. Sicilia. Associating clinical archetypes through umls metathesaurus term clusters. *Journal of Medical Systems*, pages 1–10, 2010. 10.1007/s10916-010-9586-9.

[8] R. Qamar, J. Kola, and A. Rector. Unambiguous data modeling to ensure higher accuracy term binding to clinical terminologies. In *AMIA Annual Symposium Proceedings.* American Medical Informatics Association, 2007.

[9] R. Qamar and A. Rector. MoST: A System to Semantically Map Clinical Model Data to SNOMED-CT. In *Semantic Mining Conference on SNOMED-CT*, pages 38–43, 2006.

[10] E. Sundvall, R. Qamar, M. Nystrom, M. Forss, H. Petersson, D. Karlsson, H. Ahlfeldt, and A. Rector. Integration of tools for binding archetypes to snomed ct. *BMC Medical Informatics and Decision Making*, 8(Suppl 1):S7, 2008.

[11] S. Yu, D. Berry, and J. Bisbal. An Investigation of Semantic Links to Archetypes in an External Clinical Terminology through the Construction of Terminological "Shadows". In *Proceedings of International Association for Development of the Information Society e-health 2010*, pages 9–17, 2010.