Method for the Mapping between Health Terminologies aiming Systems Interoperability

Thiago Fernandes de Freitas Dias
Inter Bioengineering Postgraduate Program, School of Engineering of São Carlos/Faculty of Medicine of Ribeirão Preto/Institute of Chemistry
University of São Paulo - USP
Ribeirão Preto, São Paulo, Brazil
thiagoffdias@gmail.com

Domingos Alves
Department of Social Medicine, Faculty of Medicine of Ribeirão Preto
University of São Paulo - USP
Ribeirão Preto, São Paulo, Brazil

Joaquim Cezar Felipe
Department of Computation and Mathematics, Faculty of Philosophy, Sciences and Languages of Ribeirão Preto
University of São Paulo – USP
Ribeirão Preto, São Paulo, Brazil

***Abstract*—The translation or mapping between terminologies in healthcare is one of the ways to achieve semantic interoperability between information systems. Accordingly, we propose a method to assist the translation of terminologies based on the association rules mining in integrated databases containing data encoded with two different terminologies. This method also uses text search (string matching) between terms. The extracted rules proved the correct translation of some terms and when a valid rule could not be extracted, textual search proved to be a good resource. Further work will be undertaken to quantify the efficiency of the method through expert analysis.**

Keywords—Health Information Systems; Health Terminologies; Terminology Mapping; Association Rules;

I. Introduction

The exchange of data between health information systems in a semantic manner requires that the shared information between the systems are understood at the level of formal definition of concepts [1], i.e., the information must have the same meaning in all systems.

The prerequisites for semantic interoperability between systems mainly include compliance with standards, reference models and domain specification, use of terminologies and specific domain ontologies [2].

These methodologies can be used to support semantic interoperability of health information systems, through the same clinical terminology sharing or use of mappings between terminologies and encodings. They are also able to capture clinical concepts and terms, preserving their semantic meaning and still pose health knowledge [3]. These capabilities are desired in natural language processing, indexing medical records and literature as well as in decision support systems.

Clinical terminologies are a standardized set of terms to record events and health interventions detailed enough to support the health care process, decision support, clinical research and improving the quality of health care, so that the reporting is done in an organized and consistent manner, ensuring that a term has only one meaning. Accordingly, two terminologies that represent the same context can assume relations allowing that the terms can be semantically consistent.

Some key terminologies and health standards used internationally for the encoding of terms and concepts in health are SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms), ICD-CM (International Classification of Diseases, Clinical Modifications), ICD (International Classification of Diseases), LOINC (Logical Observations Identifiers, Names and Codes), International NANDA (North American Nursing Diagnosis Association) and UMLS (Unified Medical Language System). In Brazil we can find the Procedures Table of SUS (Unified Health System) and CBHPM (Brazilian Hierarchical Medical Procedures Coding) from Brazilian Medical Association. Regarding supplementary healthcare, we have the initiative of the National Health Agency – in order to better regulate and standardize the information exchanged in the supplementary healthcare in Brazil – that created the standard TISS (Exchange of Information on Supplementary Healthcare), together with TUSS terminology (Unified Terminology of Supplementary Healthcare).

Given this varied terminology context, is important to create translations and mappings between health terminologies, able to relate two different terminologies, however of the same domain, so it is necessary that the terms represent the same concept [4], [5]. These translations and mappings are performed through the help of domain experts, manually, or by some computational method, facilitating such work, which shows extremely costly due to the large extent that terminologies in health present.

Bouhaddou et al. [6] described a framework developed for translation between terminologies of local American health agencies, using standard terminologies (RxNorm, UMLS, for example) as mediators in this translation. Cimino et al. [7] also presented the translation of two terminologies, ICD-9-CM to MeSH using the UMLS as a mediator.

978-1-4799-6644-8/14/$31.00 ©2014 IEEE

Rocha and authors, [8] presented an automatic translation between ICD-9-CM and MeSH, through the representation of these terminologies in a standardized way, that it was possible to represent a semantic description of terms. In this work, they used a semantic network to support the translation.

The work of Merabti et al. [9] described how the mapping was done between the Orphanet terminology and ICD10 and between MeSH and Orphanet. They used two approaches, the first manual and the second in an automatic way, by means of string's matching of Orphanet and Mesh terms. In another study [10], the authors presented another translation between terminologies similarly to the previous work.

Text processing and terms' searching in terminologies are some of the major methods used for automatic translations between terminologies. These methods can be further enhanced with the use of intermediary tools, such as ontologies, thesauri, among others, being the most common the use of UMLS. However, when using this method, it is necessary that both terminologies and helping tool have being in the same language.

Often it is not possible to employ this approach, since we do not always find ontologies or UMLS, in the language we want. This is the case of the Portuguese language, which does not yet have a translation of the main terminologies used as mediators, e.g.: UMLS SOMED-CT and MeSH.

Thus, the aim of this work is to propose an alternative method of translation between terminologies in health that is language independent and does not need any mediator assistance tool. This translation method is based on association rules extracted from an integrated databases containing information encoded by two different terminologies (which you want to translate), so that each record has been encoded by both terminologies. In this method, we also count with the participation of domain experts that will make the analysis of extracted rules, establishing the correct correlation between the terminologies.

II. Association Rules

One of the areas of Data Mining which has great relevance is the area facing Association Rules Mining, which has been developed to identify in data bases, strongly associated relationships, that have high frequency and strong correlation [11].

Association rules are denoted by X → Y (X implies Y), i.e. an item or set of items X imply a certain item (or set of items) Y, in the same transaction. According to Zhang et al. [11], I = {i1, i2, i3, ... im} is a set of items and Ai = v is an item where v is the value of attribute Ai in a R = {A1, A2, A3, ..., An} relationship. By the definition of itemset, we have: X is an itemset I, if X is a subset of I. The set of Ai = v is an itemset of a relationship R = {PID, A1, A2, A3, ..., an}, where PID is a key. A transaction t is an instance of the relation R, and a transactional database (D) is a set of transactions T, i.e. D = {ti, ti +1, ..., ti + n} and t = {tid t-itemset}, where each transaction t is composed of a key tid and a t-itemset.

From these relations we can extract a support measure of an itemset. An X itemset, in a transactional database D, have support denoted by supp(X) (1), that represents the ratio of transactions in D that contain X, namely:

$$supp(X) = \frac{|X(t)|}{|D|} \tag{1}$$

Where, X(t) = {t in D | t contain X}.

An itemset X is considered frequent if it has a support greater than a certain minimum support (minsupp), defined by an expert.

Taking into account these definitions, we can define the association rule X → Y as the itemset X implies the itemset Y if X ∩ Y = ∅.

Association rules have, in addition to a measure of support (2), also a measure of confidence (3), (4). The support of an association rule X → Y is [11]:

$$supp(X \to Y) = supp(X \cup Y) \tag{2}$$

The confidence of a rule X→Y is:

$$conf(X \to Y) = \frac{|(X \cup Y)(t)|}{|X(t)|} \tag{3}$$

Or,

$$conf(X \to Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{4}$$

Thus we have that support means the frequency with which a certain X occurs with Y in the database, and confidence is the force with which this binding occurs.

The extraction of these rules is performed by algorithms, which analyze transactional databases, providing as output a set of valid frequent rules, i.e. having support and confidence greater than or equal to a minimum support and confidence. The most commonly used algorithm for this purpose is the Apriori Algorithm [11], [12].

III. Methods

The design of this proposed method consists to perform the integration of two separate databases containing the same information previously encoded by different terminologies, with the goal of building a transactional database, from which it will be possible to conduct association rules mining for the extraction of knowledge, which in this case represent the translation between terminologies. Each extracted rule possesses the measures of support and confidence given by the algorithm.

According to every rule extracted and their measurements, it will be possible to assess how such correlation is correct, and suggest that the relationship is even a valid translation for the terms of these terminologies or not.

The results obtained through the mining of association rules will be presented to domain experts for analysis. Along with these results also will be presented results generated by textual search (strings matching), for the same terms. This text search was implemented using Lucene[1] library, a Java library that allows among other things, the text indexing and searches. Thus, the terms of one terminology (reference terminology) were indexed and subsequently queried with the terms of another (target terminology), returning the terms matching. To perform the indexing and query implemented by Lucene, the classes "BrazilianAnalyzer" and "QueryParser" were used respectively.

Therefore, a reference code will be displayed to the expert, who must translate it to the target terminology. The translation will be aided by the results obtained with the rules mining and also from textual search. By this way, the experts will decide whether the evaluated rule is a valid translation or not, taking into account their prior knowledge, the measurements of each rule and also the discrepancies between the rule and the result presented in textual search.

During the expert interaction with the system, he has four options to select the mapping, as it deems proper, between the reference code and the target terminology code:

1. Association rules: option provided by the association rules mining, which presents a set of associations between the reference code and possible codes of the target terminology, with their support and confidence measures. This option cannot be provided for a given reference code if it was not extracted an association rule that has minimum support and confidence. The support and confidence attributed for the association rules mining was 1% for both measures, because we want to extract the most number of associations.
2. Textual Search: option provided by text search from the similarity between the words that make up the textual description of the reference code with the terms of target codes. This option can not be provided for a given reference code if the textual search has found no target code whose textual description has terms similar to those of the reference code.
3. Suggestion of the user: the user enters a target code that knows a priori as the most suitable for the reference code.
4. No code: the user chooses not to select any code, by not agreeing with the suggestions and do not know which target code to assign to the reference code.

According to these options, we can see that only the first two are for some aided method (association rules or textual search) and the other two options do not.

The validation of this method will be carried out analyzing the interaction of specialists with a system, specifically designed to assist in the translation between the terminologies. In this case the terminologies will be the Unified Terminology for Supplementary Healthcare (TUSS) and International Classification of Diseases - 9th Revision – Clinical Modification (ICD-9-CM), in Portuguese version. These terminologies were chosen for validation of the methodology because of the importance they have in the national and international stage, respectively, related to the coding of surgical procedures for payment and reimbursement. In addition, this translation will allow the comparison of data generated in Brazil with data from other countries like the United States.

In this validation process, will be used a database of hospital admissions in the region of Ribeirão Preto – SP - Brazil. These data were obtained from the Processing Center of Hospital Data (CPDH) from Faculty of Medicine of Ribeirão Preto (FMRP-USP), available in two databases, whose surgical procedures are coded following the ICD -9-CM and the TUSS. These databases were integrated with the aid of Pentaho Data Integration Tool[2], allowing the construction of a transactional database with different sources of coding in a single transaction.

The integrated database served as input for Apriori algorithm, implemented by Christian Borgelt [13], responsible for carrying out mining of association rules. From this point, the rules generated were stored in a MySQL[3] relational database, acting as the basis for all the developed method. The implementation of this method also included the creation of a Web system using Java EE[4] technology, to facilitate access and use by specialists.

IV. Results

From experiments performed with a set of voluntary medical experts, we have extracted some interesting results regarding the mined rules. In the example showed in Fig. 1, we can see that two rules, or mapping options were extracted for TUSS code 31002129: ICD-9-CM 4399 (support 0.074, confidence 85.71) and 4469 (support 0,074; confidence 14,28). Analyzing the support and confidence of these rules, the algorithm suggests that the correct mapping for the presented TUSS code will be ICD-9-CM 4399 (i.e. TUSS 31002129 to ICD-9-CM 4399). According to these results displayed by the system, it is easy and quick to the expert defines the correct mapping because the codes have been previously suggested by the system.

We also note that there is not always a rule extracted as suggested mapping, e.g. for TUSS code 31002196 (Fig. 2). In this example, it is presented to the expert only suggestion found by text search (string matching), which can be accepted or rejected.

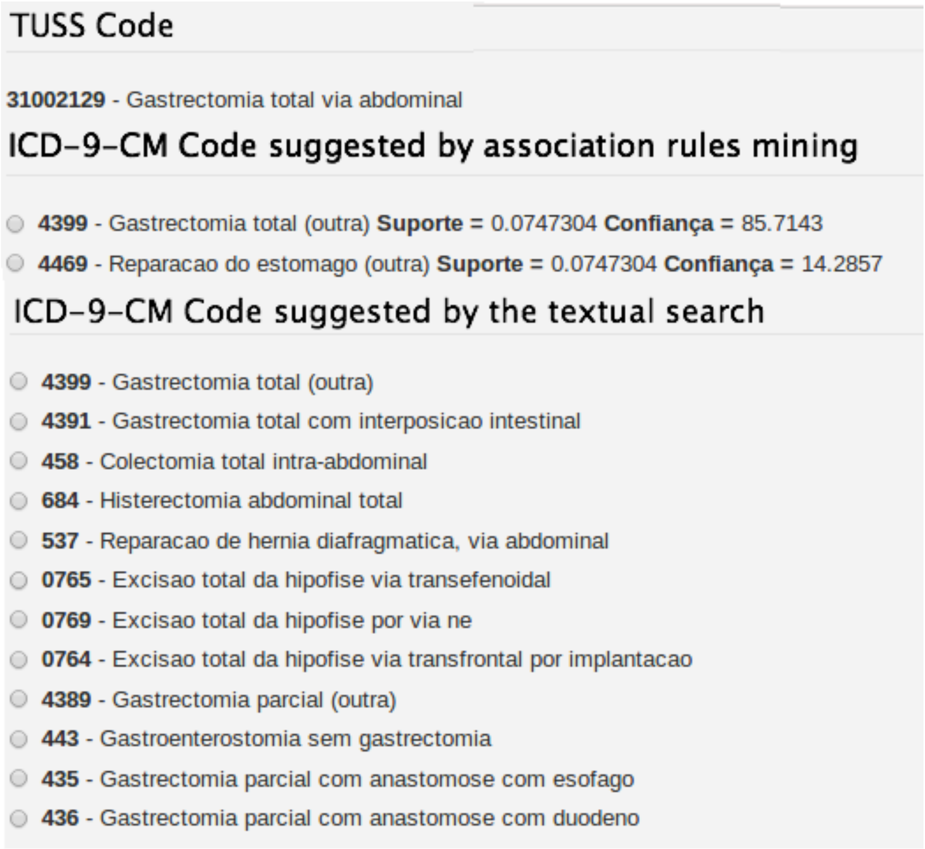
There is also the possibility of any suggestion not be displayed to the expert because it is not always possible to extract a valid rule or find a result in text search, when this occurs, the system offers the expert the possibility to enter manually the ICD-9-CM code it deems most appropriate.

[1] http://lucene.apache.org/core/4_7_1/index.html

[2] http://community.pentaho.com/projects/data-integration/

[3] http://www.mysql.com

[4] http://www.oracle.com/technetwork/java/javaee/overview/index.html



TUSS Code

31002129 - Gastrectomia total via abdominal

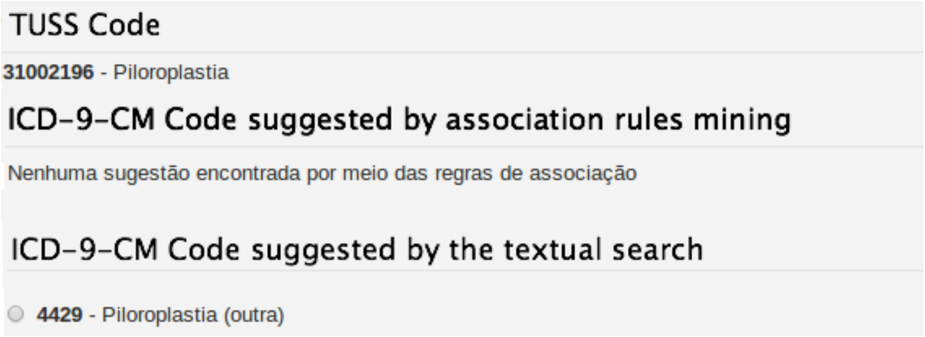
ICD-9-CM Code suggested by association rules mining

○ 4399 - Gastrectomia total (outra) Suporte = 0.0747304 Confiança = 85.7143
○ 4469 - Reparacao do estomago (outra) Suporte = 0.0747304 Confiança = 14.2857

ICD-9-CM Code suggested by the textual search

○ 4399 - Gastrectomia total (outra)
○ 4391 - Gastrectomia total com interposicao intestinal
○ 458 - Colectomia total intra-abdominal
○ 684 - Histerectomia abdominal total
○ 537 - Reparacao de hernia diafragmatica, via abdominal
○ 0765 - Excisao total da hipofise via transefenoidal
○ 0769 - Excisao total da hipofise por via ne
○ 0764 - Excisao total da hipofise via transfrontal por implantacao
○ 4389 - Gastrectomia parcial (outra)
○ 443 - Gastroenterostomia sem gastrectomia
○ 435 - Gastrectomia parcial com anastomose com esofago
○ 436 - Gastrectomia parcial com anastomose com duodeno

Fig. 1 - Suggested mapping presented by the system, by means of association rules mined and also for text search (strings matching)



TUSS Code

31002196 - Piloroplastia

ICD-9-CM Code suggested by association rules mining

Nenhuma sugestão encontrada por meio das regras de associação

ICD-9-CM Code suggested by the textual search

○ 4429 - Piloroplastia (outra)

Fig. 2 - Suggested mapping presented by the system, by textual search (strings matching)

There is also the possibility of any suggestion not be displayed to the expert because it is not always possible to extract a valid rule or find a result in text search, when this occurs, the system offers the expert the possibility to enter manually the ICD-9-CM code it deems most appropriate.

Tables 1 and 2 present statistics of results obtained by the interaction of experts to a set of mappings performed taking the TUSS terminology as reference and ICD -9 - CM as target terminology. In Table 1 we can see that when the two aided options (association rules ant textual search) are presented to the expert judgment, in 36.8 % of the mappings was chosen the suggestion made by the association rules, against 29.5 % of the choices by the suggestion of textual search, indicating that the suggestions by the association rules helped experts more than suggestions for text searching.

TABLE 1. Methods selected by experts when were simultaneously presented options of the two aided methods (association and textual search).

Selected method	Total of mappings	Percentage (%)
Association rules	35	36.8
Textual search	28	29.4
User's suggestion	6	6.3
No code	26	27.3
Total	95	100

If we consider the situation where the expert did not opt for any kind of aid, i.e. not assigned any ICD-9-CM code for TUSS code presented, or else gave another code that was not present in the suggestions, according to Table 2 we observe that in most cases (75 %) the experts refused the assistance offered by text searching, as just assigning another code (if any) even if there is a list of codes offered as a suggestion for text searching.

TABLE 2. Method displayed to experts when they chose no code or suggested some.

Displayed method	Total of mappings	Percentage (%)
Association rules + textual search	32	16.4
Association rules	5	2.5
Textual search	147	75.3
None	11	5.6
Total	195	100

V. Discussion and Conclusion

Taking into account the results of the mappings obtained so far, we can say that our method helps experts in health terminology translations tasks, because most of the time which was presented suggestions codes through association rules and textual search, the experts chosen one of the two options and the option of association rules were chose more frequently. So, experts are responsible just to accept or reject a suggestion, without needing to manually search in terminology, for the code that would be the correct.

Evaluating the results only when the expert has not opted for any code, we can see that the suggestions offered by textual search did not help experts, because they refused mostly all the time these suggestions.

The focus of this method is to provide suggestions through the extracted association rules, because they have the support and confidence measures that greatly assist the decision of an expert, who cannot know the terminology that is translating, but must have domain in the specific area of these terminologies' application. Therefore our method allows the translation of any terminology in health (or other areas), if there is a database that brings together the same information previously encoded as in one terminology as another.

In general, the method has shown good results pointing to its effectiveness. As future work we pretend to obtaining a volume of maps made by experts that is large enough to proceed with the statistical analysis of the process, quantifying how much the proposed method has helped in the decision by a code or other, comparing amount of times the expert chosen by the suggestion code provided by the association rules or by text search (Lucene), establishing a measure attesting to the effectiveness of the method. This step of acquiring the mappings performed by specialist is already in progress.

References

[1] I. O. for S. ISO, "ISO/TC 215 Technical Report: Electronic Health Record Definition, Scope, and Context." 2003.

[2] D. M. Lopez and B. G. M. E. Blobel, "A development framework for semantically interoperable health information systems.," *Int. J. Med. Inform.*, vol. 78, no. 2, pp. 83–103, Mar. 2009.

[3] K. M. Coonan, "Medical informatics standards applicable to emergency department information systems: making sense of the jumble.," *Acad. Emerg. Med.*, vol. 11, no. 11, pp. 1198–205, Nov. 2004.

[4] GEM, "Procedure Code Set General Equivalence Mappings ICD-10-PCS to ICD-9-CM and ICD-9-CM to ICD-10-PCS 2007 Version Documentation and User's Guide." pp. 1–28, 2007.

[5] I. H. T. S. D. O. IHTSDO, "Mapping SNOMED CT to ICD-10 Technical Specifications." 2012.

[6] O. Bouhaddou, P. Warnekar, F. Parrish, N. Do, J. Mandel, J. Kilbourne, and M. J. Lincoln, "Exchange of Computable Patient Data between the Department of Veterans Affairs (VA) and the Department of Defense (DoD): Terminology Mediation Strategy," *J. Am. Med. Informatics Assoc.*, vol. 15, no. 2, pp. 174–183, 2008.

[7] J. J. Cimino, S. B. Johnson, P. Peng, and A. Aguirre, "From ICD9-CM to MeSH using the UMLS: a how-to guide.," *Proc. Annu. Symp. Comput. Appl. Med. Care*, pp. 730–4, Jan. 1993.

[8] R. A. Rocha, B. H. Rocha, and S. M. Huff, "Automated translation between medical vocabularies using a frame-based interlingua.," *Proc. Annu. Symp. Comput. Appl. Med. Care*, pp. 690–4, Jan. 1993.

[9] T. Merabti, M. Joubert, T. Lecroq, A. Rath, and S. J. Darmoni, "Mapping biomedical terminologies using natural language processing tools and UMLS: Mapping the Orphanet thesaurus to the MeSH," *IRBM*, vol. 31, no. 4, pp. 221–225, 2010.

[10] T. Merabti, P. Massari, M. Joubert, E. Sadou, T. Lecroq, H. Abdoune, J.-M. Rodrigues, and S. J. Darmoni, "An automated approach to map a French terminology to UMLS.," *Stud. Health Technol. Inform.*, vol. 160, no. Pt 2, pp. 1040–4, Jan. 2010.

[11] C. Zhang and S. Zhang, *Association rule mining: models and algorithms*. Springer-Verlag, 2002.

[12] R. Elmasri and S. B. Navathe, *Fundamentals of Database Systems*. 2011.

[13] C. Borgelt, "Frequent item set mining," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 6, pp. 437–456, Nov. 2012.