

“Crowdsourcing” as a Means to Identify SNOMED CT Subsets – an Initial Approach

Dave Parry

*School of Computing and Mathematical Sciences
Auckland University of Technology, Auckland, New Zealand
Dave.parry@aut.ac.nz*

Abstract

Crowdsourcing is a technique that uses the internet to allow large numbers of people to contribute small amounts of time and effort to a research project. Selection of SNOMED CT subsets and confirmation of algorithmic coding seems particularly appropriate use of this approach. The aim of this project is to confirm that the use of a fuzzy ontology based medical coding systems is practical and effective. This paper describes the development of a prototype system that uses fuzzy logic to characterise the membership of a concept within a subset. Further development will require public access to such an engine to allow learning of relations and confirmation of their utility. Effectiveness measures including ease of use and quality of coding will confirm or refute the validity of this approach.

1. Introduction

SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) (<http://www.ihtsdo.org/snomed-ct/>) is a very large clinical vocabulary that is increasingly being adopted as a means of standardised recording of medical information. SNOMED CT is free to use within New Zealand and has superseded READ codes. There are over 300,000 concepts in the latest release and over 1 million relationships between these concepts. A number of support files are included with the release which allow mapping from fragments, the identification of synonyms and the use of modifiers. One of the major advantages of the recent SNOMED releases is that the structured nature of the vocabulary allows an ability to map from SNOMED to ICD 9 and files are provided for this purpose. Generally the co-occurrence of a number of SNOMED concepts can imply an ICD diagnosis. This can simplify the problem of automating discharge summaries and entering data into decision support systems, as coding directly from free text to definitive ICD codes is often difficult. However because of the very wide range of concepts, and the rate of change of those concepts [1], dealing with the entire set of SNOMED concepts is difficult. In particular when clinicians use SNOMED browsers to code clinical documents, experiments have shown their performance to be disappointing [2]. Given the very large amounts of textual information generated in healthcare, a model of coding similar to (Figure 1) may give the best combination of freedom of data entry and accuracy and reusability of data.

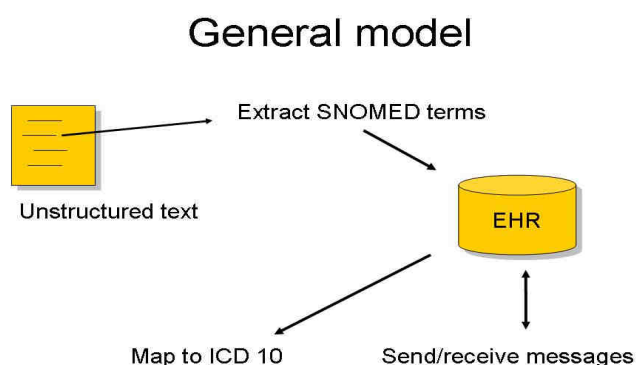


Figure 1 - Automated coding via SNOMED

2. Literature review

2.1. Crowdsourcing

Crowdsourcing [3] refers to the use of small amounts of time and effort from a large number of individuals to solve large problems which may or may not benefit those people directly. The Internet has made such efforts much more attractive as large numbers of people use the web as a communication channel for much of their work and leisure. Projects like the search for extraterrestrial intelligence (SETI@home) preceded full crowdsourcing applications [4]. However while SETI@home uses otherwise idle computer power, crowdsourcing uses the brainpower of people. There are a number of applications that have demonstrated the attractiveness of collaborative projects on the Internet with a growth in the use of social or collaborative bookmarking [5] Examples include del.icio.us (<http://del.icio.us/>) and reddit (<http://reddit.com>). On these sites, users are encouraged to share bookmarks and hence create a user-generated index of webpages along with ratings and association with shared index terms. Similar systems are used within on-line retailers such as AMAZON where reviews of books provide additional index terms for searches of the website as well as recommendations.

More formal collaborative projects are already relatively common on the web – the most famous example being Wikipedia. In the area of semantic research the OpenMind Commonsense project [6], has captured more than 70,000 valid sentences, that allow a representation of the world which may be useful for artificial intelligence research. One of the benefits of such approaches is that people can identify their own areas of expertise and the limits of their knowledge. Although syntactic rules are often general, the semantics of particular examples are often complex and difficult to represent in a general form. There is a vast potential to learn from users, the most fruitful approaches may involve correcting. By allowing the users to gain from their activities – by getting better matches and coding – the problem of malicious or unintelligent information coming from users should be reduced. This approach reverse the “broadcasting” or publication model where one person generates knowledge and many use it – in these approaches the many generate useful content for the many or the few [7].

A number of areas in medical informatics and coding may benefit from this approach. By using fuzzy logic this project attempts to assist with the identification of relevant concepts for a particular domain – in this case women’s health ultrasound.

2.2. SNOMED subsets and coding

Coding medical unstructured text into such coding systems as ICD10 is difficult and demanding, the SNOMED CT vocabulary [8] allows mapping from concepts to other coding systems and is becoming a standard in many countries. There have been some promising attempts to automate the process of transforming free text to SNOMED for example [9], but issues arise with the sheer scale of the task, One particular issue is that many areas of clinical endeavour require particular subsets only in order to allow precise and unambiguous selection of concepts and that abbreviations may be different for different areas [10]. This is particularly important in systems that involve the selection of codes by people who may work across areas, and where a logic test i.e. that the code “makes sense” may not be possible. As previously mentioned, browser based coding by clinical staff is often poorly performed, and this may be partially due to the very large range of concepts. One way around this is creation of smaller code sets, for example Read codes [11], but these bring their own problems, lacking precision and coverage in some cases.

Code subsets are attractive because they remain within the hierarchy of the overall coding system, and continue to use the richness of relations and modifiers, while not overwhelming the user. If the subset is inadequate, it can be expanded. By retaining whole branches of the hierarchy within the subset, transitive relations to route terms can be retained. SNOMED CT supports the creation of subsets, and procedures can be put in place to maintain them. Subsets also allow the disambiguation of terms, where a term may be related to multiple concepts, but only one of those concepts lies within the selected subset. For practical; system design, very large vocabularies are difficult to represent using combinations of picklists and cascading menus.

2.3. Fuzzification of subsets

One issue that arises when selecting subsets of concepts is the problem of defining the edge of a domain that is which concepts should be included and which excluded. This is particularly acute in areas that may cross traditional speciality boundaries. Women’s health ultrasound (WHU) is a case in point, where imaging reports may refer to the woman and/or the fetus if she is pregnant. Potentially very large “subsets” may then exist and there is the possibility of ambiguity as to which concept is associated with each term.

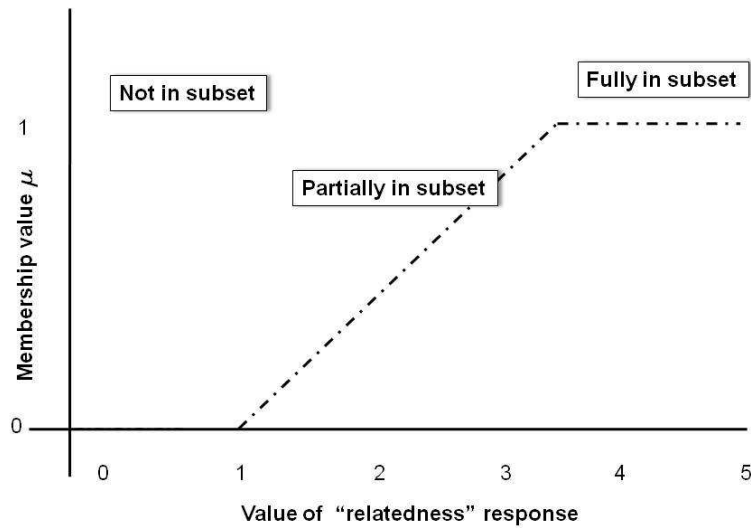


Figure 2 - Degree of membership of a concept

A well known method of dealing with such problems is the use of fuzzy logic [12]. In a fuzzy logic representation the degree of membership μ is used where $0 < \mu < 1$, and μ depends on the membership function of a fuzzy modifier such as “strongly”, “partially”, “somewhat”, “slightly” of a particular concept in a particular subset – see Figure 2. There are other approaches, for example that described in [13], where the fuzzy relations are seen as encoding the similarity between concepts.

In the figure below (Figure 3) Concepts A and B are members of both subsets – concepts related to WHU to some degree, but not wholly part of either. Examples of this may be concepts such as Doppler ultrasound scan of umbilical artery (427623005) which lies clearly within an WHU subset, whereas Ultrasound vascular - Doppler effect (241446003) may or may not be associated with a WHU procedure. This becomes especially important if the system is being used by different groups of users who may work across subsets, where the degree of membership of each subset can be used to produce potential relevance scores.

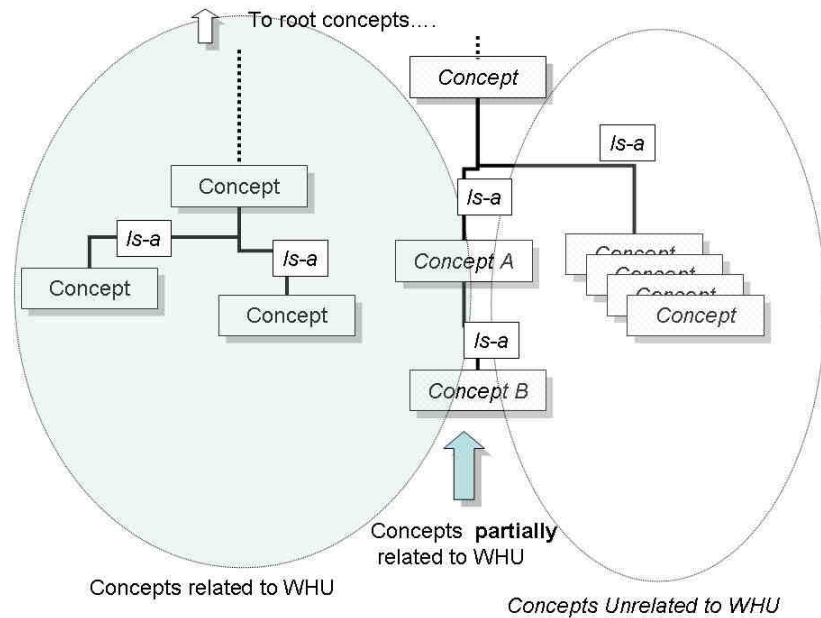


Figure 3 - Example of Fuzzy borders between subsets

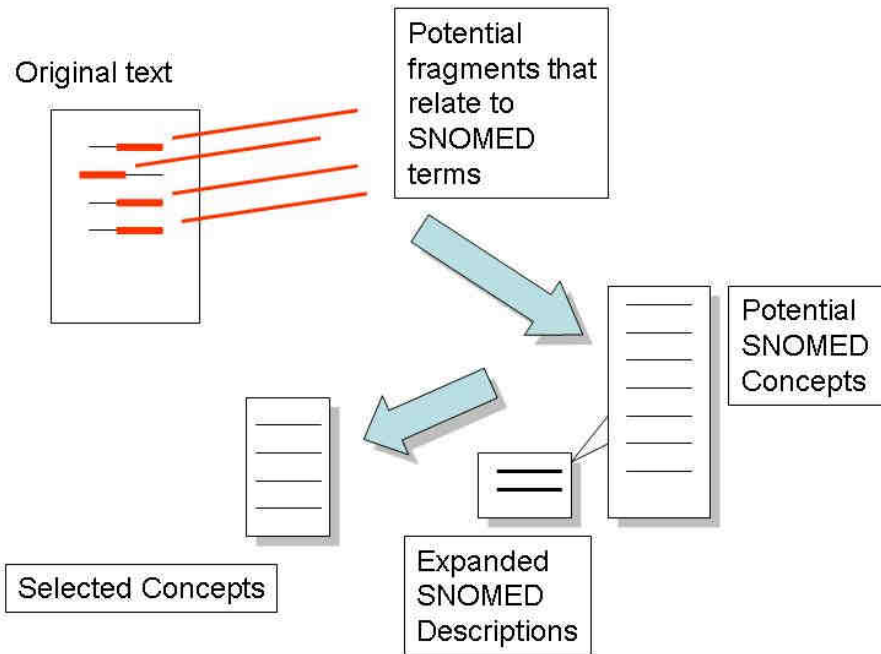


Figure 4 - System Architecture

3. Materials and Methods

3.1. System design

The overall system architecture is shown in Figure 4. Users are presented with one of a number of fictional ultrasound reports, or they can enter their own into the parser. Using the word key index, potential SNOMED terms are highlighted in the text box and a list of potential concepts created, ordered by membership of the subset.. When the user selects a concept the expanded description will be displayed. The user then selects the appropriate concept. The user will also rate the degree of membership of each concept to the subset.

3.2. Prototype

The current prototype has been developed using ASP.net and SQL server 2005 Database (Microsoft). SNOMED files were imported from the 2009 release supplied via NZHIS. The main tables used are the concept, description, relationships and wordkey index. Screen shots are shown below (Figures 5 and 6).

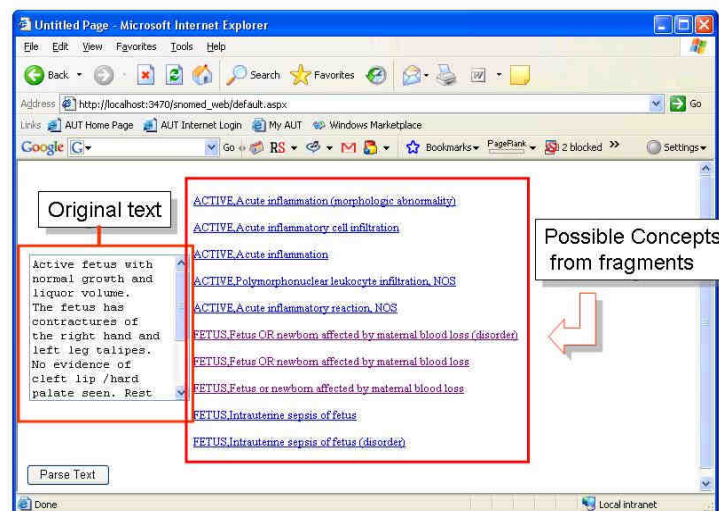


Figure 5 - Initial text entry and parsing screen

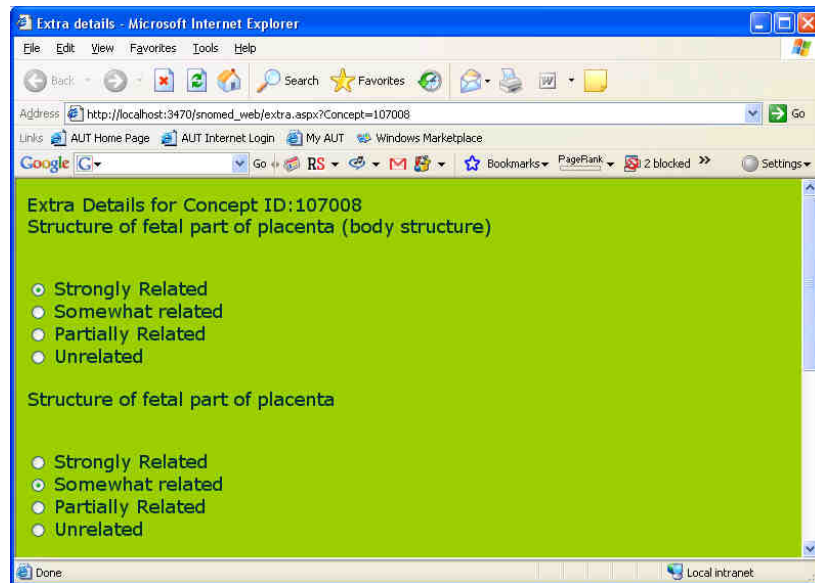


Figure 6 - Concept rating screen

3.3. Current research issues

One of the major areas of work is the so-called “cold start” problem. This occurs because until a large number of concepts have received ratings the system is unlikely to rank the concepts accurately as they will all be unseen. In order to avoid this a basic subset of WHU-related concepts was developed by inspection of the hierarchy, and these have been given a membership value of 1, whereas other terms have been given an initial value of 0.2. These values were chosen to discriminate against the unexpected values but also allow for modification of the original set.

The learning algorithm used to identify the degree of membership of concepts in the subset is based around a simple algorithm. This includes the number of documents that include the potential concept along with the initial membership value for that term. The updating of the fuzzy ontology membership value is weighted by an extremely simple algorithm where the new membership (μ_{New}) is determined by the old membership (μ_{Old}) the membership calculated for this query (μ_i), and the number of queries that have confirmed the intended meaning of this term (Q_{Hist}).

$$\mu_{New} = \mu_{Old} \pm (\sqrt{(\mu_i - \mu_{Old})^2} / Q_{Hist})$$

This learning function is designed to rapidly reflect major changes, but to prevent major oscillations in value over time.

4. Discussion

Using crowdsourcing approaches allows us to implement the idea that “All of us are smarter than any of us”. It may provide a valuable middle way between the fully automatic approach used by many information retrieval processes and the need to teach and audit the coding of large numbers of clinical workers. The membership map, and the variation in membership produced by individuals produced by this sort of work may also provide information that is useful for the development of coding systems.

The drawbacks of crowdsourcing include the danger of bias or ignorance on the part of the participants adversely affecting the outcomes. One way round this is to enlist the cooperation of professional bodies, for example the medical colleges and organisations such as the World Federation for Ultrasound in Medicine and Biology.

5. Future work

Currently the system is only a prototype but the first phase of the research project will use the “build, test & refine” [14] software development prototyping methodology to implement the technical equipment and the database application.. During this first phase continuous usability improvements will be incorporated using cognitive walkthrough [15] and

heuristic evaluation [16] by the development team and colleagues. The initial crisp ontology will be modified from SNOMED, but learning will take place on the documents (anonymised ultrasound reports) present in the website.

The second stage will involve internal testing of the software and refinement of the interface, along with learning of relations from user involvement. Internal testing will require ethical approval, as test users will be observed using the system and their responses noted. This will be a very similar process to that described in a previous version of the system, that was not browser based described in [17].

Once the software is ready for public release, the third phase will involve true crowdsourcing applications. The focus will change to the use of the system to confirm or reject the mapping of stems to concepts in a set of plausible but fictitious ultrasound reports. Clinical staff from the ultrasound and women's health community will be invited to use the system and comment. They will also be able to cut and paste their own documents into the system to attempt to extract SNOMED terms. No data pasted onto the site will be retained and no information about the users retained.

During the trial the users queries and activity will be logged, and a sample of users will be asked to fill in a pop-up satisfaction survey. The degree of satisfaction with each concept mapping result returned will also be recorded by means of a screen response box.

Outcome measures will include satisfaction scores from a sample of users, degree of modification of the ontology, and accuracy of pre-coded ultrasound reports.

It is planned to make the ontology and subset data freely available to interested parties, and encourage collaboration with teams from different areas of clinical practice. A coding portal, to allow conversion of free text into SNOMED concepts using a parser based on this data is also planned

6. Acknowledgements

Particular thanks to the women's health ultrasound department at Auckland District Health Board, especially Kathy Dryden, Chief Sonographer. Anyone using SNOMED CT in New Zealand relies on the work of NZHIS and in particular Ted Cizadlo and his team.

7. References

- [1] Spackman K. Rates of Change in a Large Clinical Terminology: Three Years Experience with SNOMED Clinical Terms. *AMIA Annu Symp Proc*; 2005; 2005. p. 714–8.
- [2] Michael F. Chiang JCH, Alexander C. Yu, Daniel S. Casper, James J. Cimino, and Justin Starren. Reliability of SNOMED-CT Coding by Three Physicians using Two Terminology Browsers *AMIA Annu Symp Proc*; 2006; 2006. p. 131–5.
- [3] Aniket K, Ed HC, Bongwon S. Crowdsourcing user studies with Mechanical Turk. *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. Florence, Italy: ACM 2008.
- [4] Anderson DP, Cobb J, Korpela E, Lebofsky M, Werthimer D. SETI@home: an experiment in public-resource computing. *Commun ACM*. 2002;45(11):56-61.
- [5] Yanbe Y, Jatowt A, Nakamura S, Tanaka K. Can social bookmarking enhance search in the web? *Proceedings of the 2007 conference on Digital libraries*; 2007; 2007. p. 107 - 16
- [6] Singh P, Lin T, Mueller ET, Lim G, Perkins T, Zhu WL. Open Mind Common Sense: Knowledge Acquisition from the General Public. *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE : Confederated International Conferences CoopIS, DOA, and ODBASE 2002 Proceedings 2002*:1223-37.
- [7] Huberman BA. Crowdsourcing and Attention. *Computer*. 2008;41(11):103-5.
- [8] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform*. 2006;121:279-90.(121):279-90.
- [9] Patrick J, Wang Y, Budd P. An automated system for conversion of clinical notes into SNOMED clinical terminology. *Proceedings of the fifth Australasian symposium on ACSW frontiers - Volume 68*. Ballarat, Australia: Australian Computer Society, Inc. 2007.
- [10] Patrick J, Wang Y, Budd P, Rector A, Brandt S, Rogers J, et al. Developing SNOMED CT Subsets from Clinical Notes for Intensive Care Service. *Health Care and Informatics Review Online (HCIRO)*. Health Care and Informatics Review Online (HCIRO),. 2008 2008.

- [11] O'Neil MP, Read J. Read Codes Version 3: A User Led Terminology. *METHODS OF INFORMATION IN MEDICINE*. 1995; 34(1/2):187-92.
- [12] Zadeh L. Fuzzy Sets. *Journal of Information and Control*. 1965;8:338-53.
- [13] Tho QT, Hui SC, Fong ACM, Tru Hoang C. Automatic fuzzy ontology generation for semantic Web. *Knowledge and Data Engineering, IEEE Transactions on*. 2006;18(6):842-56.
- [14] Hevner AR, March ST, Park J, Ram S. Design Science in Information Systems Research. . *MIS Quarterly*. 2004;28(1):75-105.
- [15] Wharton C, Rieman J, Lewis C, Polson P. The cognitive walkthrough method: a practitioner's guide. *Usability inspection methods: John Wiley & Sons, Inc*. 1994:105-40.
- [16] Nielsen J. Heuristic evaluation. *Usability inspection methods: John Wiley & Sons, Inc*. 1994:25-62.
- [17] Parry DT. Evaluation of a fuzzy ontology based medical information system. *International journal of Health Information Systems and Informatics*. 2006;1(1):40-9.