The Cancerology Ontology: Designed to Support the Search of Evidence-based Oncology from Biomedical Literatures

Jantima Polpinij

Faculty of Informatics, Mahasarakham University, Mahasarakham 44150 Thailand
and
School of Computer Science and Software Engineering, Faculty of Informatics
University of Wollongong, Wollongong 2500 Australia
jantima.p@msu.ac.th

Abstract

This work proposes a new ontology, called the Cancerology, where it faces a problem of unclear analysis in a biomedical text processing because existing ontologies such National Cancer Institute's Thesaurus and Ontology do not offer some information relating to domain specific variations in terms that can be provided by the domain expert. This ontology is experimented through a method of text classification with retrieving the relevant cervix cancer abstracts relating to clinical trials from PubMed. The experimental results show more effectiveness for increasing the accuracy. This demonstrates that the Cancerology may be also effective for other areas of text processing and analysis, especially in the particular domain of oncology literature such as intelligent search service, text mining, and knowledge extraction.

1. Introduction

The tradition of ontology is a formal representation of explicit specification of a conceptualization with the set of concepts, relationships between concepts, and axioms about a target domain. Fundamentally, ontologies are required, because the ontology makes domain assumptions explicit, where an underlying implementation makes it possible to change these assumptions easily if the knowledge about the domain changes [1, 2]. Furthermore, the ontology's semantic data can be used to improve search accuracy for knowledge from unstructured data by understanding searcher intent and the contextual meaning of terms as they appear in the searchable data-space, whether on the Web or within a closed system, to generate more relevant results.

Therefore, the ontology not only describes the underlying concepts with their relationships and property, but also plays a major role in semantic application. The use of ontologies to support many domain analyses has been previously demonstrated [3, 4, and 5].

At present, there are several existing ontologies. However, new ontologies are always required, where the existing ontologies are not sufficient for using in some domain specific analysis. We present an example problem in the analysis of biomedical text relating to cancer literature.

Consider following sentence, "the combined simultaneous use of chemotherapy and radiotherapy can be described as *chemoradiotherapy*, *chemoradiation*, or *CT/R*". In fact, the terms of *chemoradiotherapy*, *chemoradiation*, and *CT/RT* should be interpreted as the same meaning during text processing. In contrast to normal processing, it cannot make understanding these terms in the same meaning. This has led to unclear analysis in a text processing, resulting in the accuracy of the text processing being decreased.

For another example, the associated term of '*stage IA*' represents the early stage of cancer. However, this associated term can be represented in alternative writing styles such as '*stage 1A*' and '*stage ia*'. This problem also requires ontology to support the pursuit of relevance.

Based on the problems above, these deficiencies are the motivation for this investigation. The *Cancerology* (**Cancer** Term Ont**ology**) is developed, where it is also unclear analysis in a biomedical text processing, and current ontologies such National Cancer Institute's Thesaurus and Ontology [6] do not offer some information relating to domain specific variations in terms that can be provided by the domain expert.

In this context, the entries of *Cancerology* are the term-concept mapping. The association in *Cancerology* can be conceptual and term relation harness synonym and variation of terms. Conceptual relation is to link between concepts and term relation that is to associate between terms based on meaning and variation of terms.

This paper is organized as follows. In Section 2, we show the overview about the Cancerology. Afterwards, Section 3 presents how to develop the ontology, and it shows the approach of the Cancerology and the experimental results in Section 4. Finally, conclusion is given in Section 5.

2. The Overview of Cancerology

The Cancerology is the set of cancer technical terms set including cancer histology, cancer stage, treatment modalities used, the timing of each modality, and the names of specific drugs used. Also, the Cancerology is organized according to meaning and variation of terms, so the cancer technical terms in close proximity are related

The entries of the Cancerology are the term-concept mapping. The associations in the Cancerology have conceptual and term relations which harness synonyms and variation of terms. A conceptual relation is a link between concepts while a term relation is an association between terms based on meaning and variation of terms.

The Cancerology consists of five main concepts: cancer histology, disease stage, therapeutic modalities timing, therapeutic modalities (treatment) used, and specific drugs used.

It is noted that the selection of specific drugs always depends on type of chemotheraphy used. Then, 'chemotheraphy' is a type of therapeutic modalities.

The majority of chemotherapeutic drugs for therapeutic modalities used can be divided into Alkylating Agents, Antimetabolites, Plant Alkaloids and Topoisomerase inhibitors, Antibiotics, Miscellaneous, and other Antitumour Agents.

In this context, one concept can be expressed by many terms and each term can be synonymy or variation. It can be illustrated as Figure 1.

Consider the concepts of the therapeutic modalities used and the specific drugs used in Figure 1. The concept of therapeutic modalities consists of three main technical terms: '*surgery*', '*chemotherapy*', and '*radiotherapy*', while the concept of specific drugs represent all drugs that are used for the stage of modality. It is noted if the therapeutic modality used is 'chemotherapy', this modality is concurrently used with the specific drugs. Therefore, the '*chemotherapy*' can be referred to the concept of specific drugs, and the concept of specific drugs can be referred to the '*chemotherapy*' in the concept of therapeutic modalities.

Simply speaking, one technical term can be referred to other concepts. However, it is noted that the expert domain knowledge does not permit '*radiotherapy*' to be paired with a drug such '*cisplatin*' in the same way.

```
< Therapeutic_modalities >
    <Surgery>
        <Syn> Hysterectomy, Lymphadenectomy, …</Syn>
    < /Surgery >
    < Radiotherapy >
        <Syn> RT, Radiation Therapy, …</Syn>
    < / Radiotherapy >
    < Chemotherapy >
        <Syn> CT, Chemoradiotherapy, …</Syn>
        <Types> Alkylating Agents </Types>
            <Drugs> cisplatin, carboplatin </Drugs>
        </Types>
        <Types> Alkylating Agents </Types>
            <Drugs> cisplatin, carboplatin </Drugs>
        </Types>
        ….
    < /Chemotherapy >
< /Therapeutic_modalities >
< Specific_drugs >
    <drug_name> clisplatin
      < therapeutic_modalities> Chemotherapy
        <types> Alkylating Agents </type>
      < /therapeutic_modalities>
    </drug_name>
    ….
< /Specific_drugs >
```

Figure 1. The example of two concepts: Therapeutic modalities used and Specific drugs used.

3. How to develop the Cancerology

This section describes a method of ontology development which is applied from the basic design of WordNet [7] and the step of ontology development [8]. It can be shown as Figure 2.

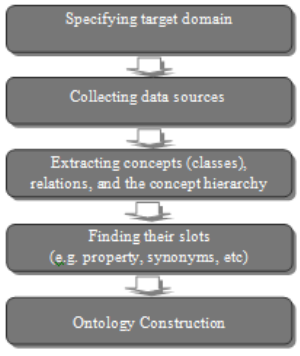

Figure 2. The Method of Ontology Development

Step 1: Specifying target domain

This ontology development method commences with determining the domain and scope of the use of ontology in order to be used as a guide for collecting a set of data sources that is necessary for developing ontology. In this context, the target domain is specified by identifying a set of keywords that are relevant to the cancer domain. A solution to obtain an initial set of keywords is to answer the questions. The questions, which are manually provided, are:

- *What is the specified domain that the ontology will cover*?
- *What is the ontology that is going to be used*?

The questions above lead to the single-word answers (as keywords), and then the keywords will be expanded to the other relevant keywords.

For example, if the answer of the question '*What is the specified domain that the ontology will cover*?' is 'cancer', this keyword 'cancer' will be expanded to other relevant keywords such as *histology*, *cancer stage*, *modality*, *specific drugs*, and so on. In addition, some of these keywords can be expanded to the others. Consider the keywords '*modality*' and '*histology*'. The keyword '*Modality*' can be expanded to *chemotherapy*, *radiotherapy*, and *surgery*, while the keyword '*histology*' can be expanded to *squamous cell carcinoma*, *adenocarcinoma*, *adenosquamous carcinoma*, and *neuroendocrine carcinoma*. Finally, these keywords are used for finding, retrieving, and collecting the relevant data sources. In the next step, it shows how all data sources are collected and organised.

Step 2: Collecting Relevant Data Sources

After obtaining a set of relevant keywords, they are used to collect the relevant data sources. This step is an important part of ontology development. In this context, there are two techniques used to collect the relevant data sources.

(1) A set of keywords is used to select and retrieve the specified relevant data collection. Two solutions are proposed following.

The first solution is to retrieve the relevant documents through using keywords and Boolean logic, and then to read the present abstract for relevance. Following that, the most relevant document that is used can be chosen as the main document, and other relevant documents can be retrieved and collected by analysing the similarity between the selected document and the candidate documents.

The second solution is to retrieve the relevant documents through using keywords and Boolean logic, and then to use the set of documents to build a model of data classification. Hereafter, the specified relevant documents are retrieved and collected by the data classification model.

(2) Another solution is to consider the reuse of existing ontologies. For example, when we develop the new ontology, the existing ontologies such as Dictionary of Cancer[1] are also used for the new ontology development in this work.

Step 3: Extracting Important Concepts (as classes) and Setting Concept Hierarchy

This step identifies important terms as ontological concepts (or classes), and organizing the concept hierarchy in ontology. In general, the concept extraction is an important step for ontology learning. The extracted concepts should be closed to objects and relationships in the domain interest. The objects can be physical or logical. Most existing ontology learning focuses on identifying concepts and taxonomic relations (e.g. *is-a*) [9]. In general, the noun is used to mention the concepts and the verb is used to mention the relationships between concepts.

This part commences with defining concepts in the domain, and setting the concepts in form of a hierarchy (or class hierarchy). In this context, the concept hierarchy is set up based on the top-down model[2] through consideration of the relationships between concepts.

After obtaining a set of concepts, the concept hierarchy is set up by considering the relationships between concepts. In this context, one or more previous ontologies can be also used to support the process of determining the concept (or class) hierarchy.

The taxonomic relations such '*is-a*' or '*kind-of*' can be used to indicate the relationship between concepts, where a class *X* is a subclass of *Y* if every instance of *Y* is also an instance of *X*. In addition, the transitive relationship role is used, where *Y* is a subclass of *X* and *Z* is a subclass of *Y,* then *Z* is a subclass of *X*. For example the transitive relationship role is illustrated, as follows.

The term '*cervix cancer*' is defined as a concept in our ontology. However, by considering it in conjunction with the cancer dictionary, it can be summarized that '*cervix cancer*' is a subclass of '*cancer*' and '*cancer*' is a subclass of '*disease*'. Therefore, transitivity of the subclass relationship

[1] http://www.cancer.gov/dictionary/

[2] A *top-down* model can be defined as the most general concepts in the domain and subsequent specialization of the concepts.

means that the class '*cervix cancer*' is a subclass of '*disease*' as well.

< *cervix cancer* > *is-a* < *cancer* >
< *cancer* > *is-a* < *disease* >
< *cervix cancer* > *is-a* < *disease* >

These can be represented in predicate logic as follows.

$\forall(x)$: *cervix_cancer*$(x) \Rightarrow$ *cancer*(x)
$\forall(x)$: *cancer*$(x) \Rightarrow$ *disease*(x)
$\forall(x)$: *cervix_cancer*$(x) \Rightarrow$ *disease*(x)

Meanwhile, '*stage*' is a description of the extent that cancer has spread. '*Squamous cell carcinoma*' is a '*cancer histology*', while '*surgery*' and '*chemotherapy*' are kinds of '*cancer modality*'. These are important aspects of information of cancer that are necessary for clinical decision-making. Some of the relationships can be represented in predicate logic.

$\forall(x)$: *surgery* $(x) \vee$ *radiotherapy* $(x) \vee$
chemotherapy $(x) \Rightarrow$ *modality* (x)
$\forall(x)$: *squamous _cell_carcinoma* (x)
$\Rightarrow$ *histology* (x)
$\forall(x)$: *stage* $(x) \Rightarrow$ *cancer*(x)
$\forall(x)$: *histology* $(x) \Rightarrow$ *cancer*(x)
$\forall(x)$: *modality* $(x) \Rightarrow$ *ancer*(x)

As a result, it can be concluded that '*cancer histology*' is a subclass of '*cancer*', '*stage*' is a subclass of '*cancer*', and '*modality*' is a subclass of '*cancer*'.

Step 4: Finding their slots

In this step, every concept is provided a slot that can have different facets such as value type (e.g. Boolean, string, number), domain and range of values, and so on. Also, the value in the slot can be the concept's meaning, and the concept's synonyms. It is noted that subclasses can inherit all slots from a super-class.

It is noted that several techniques such as a hand-crafted verification with dictionary, a similarity analysis, a probability analysis, and the Apriori Association rules [10] can be used. To select a satisfied technique depends on the desired outcome that should satisfy the need and relate to the concept.

An example of setting the concept's slot can be illustrated by the following. It is to collect a set of synonyms of a concept. A solution to this problem is to employ a simple probability analysis.

Suppose there are 10 literatures relating to oncology treatment in a clinical trial. There are four literatures using the term '*chemoradiotherapy*' to represent a combination of two modalities: '*chemotherapy*' and '*radiotherapy*'. Meanwhile, there are three literatures using the term of '*CT*', and there are three literatures using the term of '*CT/RT*' to represent a combination of '*chemotherapy*' and '*radiotherapy*'. After calculating the probability of each term, they result in 0.4, 0.3, and 0.3, respectively. These may indicate that the terms of '*chemoradiotherapy*', '*CT*', and '*CT/RT*' can be used instead of the term '*chemotherapy*'.

It is noted that these terms are also considered by a domain expert (e.g. a health processional in the oncology area) in order to obtain the correct terms that are relevant to the specific domain analysis of oncology.

Finally, the terms of '*chemoradiotherapy*', '*CT*', and '*CT/RT*' can be synonyms of the concept of '*chemotherapy*' represented in an atomic formula which is a simple form of predicate logic as

'*chemoradiotherapy*' $\vee$ '*CT*' $\vee$ '*CT/RT*'
$\Rightarrow$ '*chemotherapy*'

By obtaining all of these, we can construct the ontology which is called '*Cancerology* ' ontology.

4. The Cancerology Approach for Retrieving Clinically Relevant Oncology Literature from MEDLINE

This section describes a contribution of the Cancerology to retrieve oncology literature from MEDLINE database relevant to cervix cancer in clinical trials.

4.1 The Effectiveness of the Cancerology

The Cancerology is provided to enhance the process of cancer literature text processing. It offers following benefits.

First, it helps to identify the particular domain that users satisfy for the needs, where users need to retrieve the relevant Literature.

Second, it also helps to handle a problem of terms variation in the domain specific. This can lead to correctly analyze and interpret some cancer technical terms which are presented in different styles but they should be actually understood in the same meaning. For example the combined simultaneous use of radiotherapy and chemotherapy can be described as '*chemotherapy and radiotherapy*', '*chemoradiotherapy*', '*radiochemotherapy*', '*chemoradiation*', '*CT/RT*', or '*RT/CT*'. In the domain of text processing, the

problem of ambiguity can lead the performance of text retrieval being poor.

Third, the Cancerology may help to understand an association between concepts. For example, stage of cancer can refer to the process of treatment. If patients are in cancer stage 0, the patients can be treated by radical surgery alone. For another example, if *cisplatin* which is a cancer drug is used, it means that the clinicians use chemotherapy as the modality treatment.

4.2 The Examples of the Cancerology Approach and the Experimental Results

We show the approach of using the Cancerology through text retrieval, where the objective of this stage is to compare the different search methods between the method without the Cancerology and the method with the Cancerology. It is noted that the method of relevant literature retrieval is done through text classification.

This method commences with gathering the cervix cancer abstracts by keywords search on PubMed system, and then these abstracts are used to build a text classifier in order to use for retrieving clinically relevant cervix cancer abstracts.

The basic concept of text classification is formalized as the task of approximating the unknown target function Φ: D x $C \rightarrow \{T, F\}$ by means of a function Φ: D x $C \rightarrow \{T, F\}$ - called the classifier - where $C = \{c_1, c_2, ..., c_{|C|}\}$ is a predefined set of categories, and D is a set of documents. The documents $D = \{d_1,...,d_{|D|}\}$ to be classified are described by a vector of words $\omega = \{w_1, w_2, \ldots, w_i\}$. If $\Phi(d_i, c_j) = T$, then d_i is called a positive member of c_j, while if $\Phi(d_i, c_j) = F$, it is called a negative member of c_j.

In this context, text classifiers were implemented by two algorithms: Support Vector Machines (SVM) [11] and Naïve Bayes algorithms [12].

Before building text classifier, a collection of cervix cancer abstracts are represented in the structured '*bag of words* (BOW)'. We obtain $w = (w_1, w_2,\ldots,w_k,\ldots, w_v)$, where w_{ik} is the number of the unique words within the collection. In the BOW, a cervix cancer abstract d_i is composed of a sequence of words, with $d_i = (w_{i1}, w_{i2},\ldots, w_{ik}, \ldots, w_{iv})$, where w_{ik} is the frequency of the k-th word in an oncology document d_i. Significantly, each word is weighted by *tf-idf* (*Term Frequency - Inverse Document Frequency*) [13], which is used for providing a pre-defined set of features for exchanging information. The *tf-idf* weight of a term is the product of its *tf* weight and *idf* weight. The term frequency *tf* of term t in document d is defined as the number of times that t occurs in d. For *idf* weighting, let *df* be the document frequency of t, where t is the number of documents that contain t. It can define the *idf* of t as $idf = 1 + log\ (N/df)$.

It is noted that, the stage of word segmentation, which is the first and obligatory task in natural language processing because word is a basic unit in linguistics, relies on the Cancerology. A list of relevant keywords can be generated by using this ontology. The Cancerology is used as a dictionary.

By using the Cancerology, the process of analysis should understand that the words '*chemotherapy*', '*chemoradiotherapy*', and '*CT*' are the same meaning, although they are different for representation. Finally, the SVM and the Naives Bayes text classifiers are evaluated by the information retrieval standard [14]. Common performance measure for system evaluation can be *F-measure* (*F*), where *F-measure* is the weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is:

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{(Precision + Recall)}$$

The results of evaluation can be shown as Table 1.

Table 1. The experimental results of the SVM and the Naives Bayes (NB) text classifiers

The model without ontology		The model with ontology	
Algorithms	F-measure (%)	Algorithms	F-measure (%)
SVM	78.00	SVM	95.00
NB	72.50	NB	91.50

Consider Table 1. It can be seen that text classifiers based SVM and Naïve Bayes are more effective for retrieving the relevant cervix cancer abstracts relating to clinical trials from PubMed, after testing by *F-measure*. This is because the average accuracies can be increased. This demonstrates that the Cancerology is more effectiveness for selecting and retrieving the relevant cervical cancer abstracts from PubMed.

As above, this can indicate that the Cancerology may be effective for other areas of text processing and analysis, especially in the particular domain of oncology literature such as intelligent search service, text mining, and knowledge extraction.

5. Conclusion and Future Work

This work has proposed a new ontology, where it faces a problem of unclear analysis in a biomedical text processing. This is because existing ontologies such National Cancer Institute's Thesaurus and Ontology do not offer some information relating to domain specific variations in terms that can be provided by the domain

expert. Therefore, the Cancerology was proposed to support biomedical text processing. The background of *Cancerology* is applied from the basic design of WordNet. The entries of *Cancerology* are the term-concept mapping. The association in *Cancerology* can be conceptual and term relation harness synonym and variation of terms. Conceptual relation is to link between concepts and term relation that is to associate between terms based on meaning and variation of terms.

The Cancerology ontology is evaluated through the experiment with text classification. Our text classifiers are implemented by SVM and Naïve Bayes. Then, the results return the increasing accuracies. This may demonstrate that the Cancerology is more effectiveness for selecting and retrieving the relevant cervical cancer abstracts from PubMed.

In addition, this indicates that the Cancerology may be effective for other areas of text processing and analysis, especially in the particular domain of oncology literature such as intelligent search service, text mining, and knowledge extraction.

In future work, the Cancerology will be used in the process of the development of clinical cancer knowledge base and a semantic-search service system used as the tool to search the knowledge from clinical cancer knowledge base. This system will be used by health professionals or clinicians to support clinical decision-making.

6. Acknowledgement

I thank to Mahasarakham University Development Fund for assistance with expenses relating to travel to present the research. Special Gratitude goes to Dr. Andrew Miller, Clinical Associate Professor of Graduate School of Medicine, and University of Wollongong, who provided me with many helpful suggestions.

7. References

[1] Hruby, P., Ontology-Based Domain-Driven Design, OOPSLA Workshop on Best Practices for Model-Driven. *Software Development*, 2005

[2] Antunes, B., Seco, N. and Gomes, P., Using ontologies for software development knowledge reuse, *Proceedings of the artificial intelligence 13th Portuguese conference on Progress in artificial intelligence* (*EPIA*), 2007.

[3] Cimiano, P. and Handschuh, S. Ontology-based linguistic annotation, *2003 Proceedings of the ACL 2003 workshop on Linguistic annotation*, 2003.

[4] Vargas-Vera, M. and Motta, E. AQUA - Ontology-Based Question Answering System, *Mexican International Conference on Artificial Intelligence* (*MICAI*), 2004: 468-477.

[5] Polpinij, J. and Ghose, A.K., An Ontology-Based Sentiment Classification Methodology for Online Consumer Reviews. *Web Intelligence 2008*: 518-524, 2008.

[6] Golbeck, J., Fragoso, G., Hartel, F., Hendler, J., Oberthaler, J. and Parsia, B., National Cancer Institute's Thesaurus and Ontology, 2003 [Online] Available at http://www.globalmedicalresearch.org/research/index.php?option=com_contentandview=articleandid=254andItemid=351 [Accessed 1 March 2011].

[7] Miller, G. A., Beckwith, R., Fellbaum, C. D., Gross, D., Miller, K., WordNet: An online lexical database. *Int. J. Lexicograph.* 3, 4, pp. 235-244, 1990

[8] Noy, N.F. and McGuinness, D.L., Ontology Development 101: A Guide to Creating Your First Ontology, 2000 [Online] Available at http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html [Accessed 24 July 2010].

[9] Missikoff, M., Navigli, R., and Velardi, P., The usable ontology: An Environment for Building and Assessing a Domain Ontology, *International Symposium on Wearable Computers* (*ISWC*). pp 39-53, 2002.

[10] Agrawal, R. and Srikant, R., Fast Algorithms for Mining Association Rules, *Proceedings of the 20th International Conference on Very Large Data Bases* (*VLDB*), pp 487-499, 1994.

[11] Joachims, T., Transductive Inference for Text Classification using Support Vector Machines. In: *Proceedings of the International Conference on Machine Learning* (*ICML*), 1999.

[12] Nigram, K., Maccallum, A. K., Thrun, S. and Mitchell, T., Transductive Text Classification from Labeled and Unlabeled Document using EM. In: Machine Learning. 39(2/3). pp. 103-134, 2000.

[13] Yang, Y. and Pederson, J.O., A Comparative Study on Features selection in Text Categorization. *Proceedings of the 14th international conference on Machine Learning* (*ICML*). pp 412-420. Nashville, Tennessee, 1997.

[14] Baeza-Yates, R. and Ribeiro-Neto, B., Modern information retrieval. ACM Press, New York, 1999.