Expert Systems With Applications 114 (2018) 543–551

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Semantic term weighting for clinical texts

Ryosuke Matsuo[a,b,*], Tu Bao Ho[a,c]

[a] Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi-city, Ishikawa, Japan
[b] Faculty of Medicine, University of Miyazaki Hospital, 5200 Kiyotakecho Kihara, Miyazaki-city, Miyazaki, Japan
[c] John von Neumann Institute, VNU-HCM, Ho Chi Minh City, Vietnam

ARTICLE INFO

Article history:
Received 8 April 2018
Revised 17 July 2018
Accepted 14 August 2018
Available online 17 August 2018

Keywords:
Semantic term weighting
Two-phase framework
Causes of death ranking
Vector space model
Mortality prediction

ABSTRACT

Term weighting is an essential step to process textual data and generate input data (vector) for machine learning algorithms. In order to appropriately represent documents into computable forms for a certain task (such as text classification, clustering, sentiment analysis, recommendation and information retrieval), semantic term weighting which considers term meanings is significant for specific applications of machine learning. Two challenging issues of semantic term weighting for clinical texts are how to determine the meaning of a medical term in a given clinical text and how to give semantic weights for a huge amount of distinct terms in clinical texts. To address those challenges, this work proposes a two-phase framework for determining semantic weights of terms in clinical texts. The proposed framework derives a two-part hierarchy where each of the nodes is categories of terms. All terms in a clinical text is classified into the categories in the hierarchy and terms in the leaf nodes are assigned with the same semantic weights. Fundamentally, the deeper the hierarchy, the higher the semantic weights. The first phase classifies all terms into the categories which are commonly significant for any tasks, by using UMLS and ICD-10. These categories are organized at the first part of the hierarchy. The second phase flexibly organizes specific categories for a certain task as the second part of the hierarchy as well as the subcategories of the first part, by specific medical domain knowledge regarding the aspect under consideration. The implementation of the proposed framework for mortality prediction with semantic weights is validated by experimental comparative evaluation using the well-known EMRs database MIMIC II. The experimental results showed that the performance is considerably improved when combining frequency-based weights and semantic weights with its significant difference derived from a paired *t*-test. Although the proposed framework can be applied to only medical domain, various tasks in medical domain can be covered by the proposed framework which flexibly organizes the second part (deeper levels in the hierarchy) by specific medical knowledge regarding the aspect under consideration.

1. Introduction

In recent years, the prevalence of clinical texts such as electronic medical records (EMRs) and electronic health records (EHRs) opens new chances for developing methods to solve many significant problems in medical research regarding semantic and data integration and phenotyping (Richesson, Sun, Pathak, Kho, & Denny, 2016; Yang & Veltri, 2015). The prevalence has revealed the need for processing clinical texts. The clinical texts are narratives about patient diagnosis and treatment at hospitals. The clinical texts are considerably different from other common medical texts from literature such as medical digitalized books and research articles. Representing the clinical texts in computable forms is required for further tasks of text processing.

The vector space model (VSM) is powerful for various language processing tasks in which term weighting – giving a numerical weight to each term appearing in a document in terms of its importance for the document – plays a crucial role. Term weighting has been used in various language processing tasks, including text classification, clustering, sentiment analysis, recommendation and information retrieval. The traditional measures used in term weighting are frequency-based ones, notably TFIDF (Salton & Buckley, 1988), derived from the document frequency and the inverse document frequency of terms. TFIDF and its variants are commonly used as weights of terms in VSM. TFIDF is simple and effective, and it forms a popular base for advanced algorithms in spite of its age (Ramos, 2003).

[*] Corresponding author at: Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi-city, Ishikawa, Japan.
E-mail addresses: matsuor@jaist.ac.jp (R. Matsuo), bao@jaist.ac.jp (T.B. Ho).

https://doi.org/10.1016/j.eswa.2018.08.028
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

The term importance captured by frequency-based weighting methods does not relate to the term meanings in the domain that the term belongs to. However, there are applications that require considering the semantics of terms where either frequency-based methods may not be appropriate or semantic term weighting can additionally do better the tasks. Therefore, semantic term weighting methods have been developed aiming at assigning weights to terms in documents based on their meanings.

Ontology-based term weighting has been pursued as an approach to semantic term weighting. Ontologies systematically represent the domain knowledge in a hierarchical structure with concepts and relationships that can exist between terms (Gruber, 1993). Tar and Nyunt (2011) and Sureka and Punitha (2012) used ontologies for concept weighting by exploiting the length of words from the association between two concepts, the correlation coefficient of words and concept probability. Zakos and Verma (2006) and Sakre, Kouta, and Allam (2009) exploited four types of conceptual information in WordNet to determine the term importance. Some work demonstrated the semantic relationship of terms based on their conceptual similarity (Jing, Zhou, Ng, & Huang, 2006; Varelas, Voutsakis, Raftopoulou, Petrakis, & Milios, 2005; Zhang et al., 2008; Zhang, Jing, Hu, Ng, & Zhou, 2007). In those work, the term weight is firstly calculated through TFIDF then adjusted in accordance with the semantic similarity of other terms in the same vector. Luo, Chen, and Xiong (2011) augmented term weights based on the relevance of terms to categories in the WordNet ontology. This work proposed a general semantic term weighting schema for text categorization. In our work, although the application fields of semantic term weighting is limited, the proposed framework can be applied to various tasks in medical domain, by reorganizing the second part in the proposed framework using specific medical knowledge regarding the aspect under consideration.

In the field of medicine, medical ontologies such as UMLS or MeSH have been exploited in semantic term weighting for medical literature. Zhang et al. conducted semantic term weighting by considering the semantic relationship of terms using the MeSH ontology (Zhang et al., 2008, 2007). The medical ontology UMLS was employed to expand queries by utilizing categories such as the UMLS concept and the UMLS synonym. The method exploiting UMLS augmented the query terms from the IDF weights based on the categories (Yu & Cao, 2009). Zhu et al. utilized UMLS to augment term weights based on the selected major UMLS semantic types for TREC 2004 Genomics Ad Hoc Retrieval Task (Zhu, Xu, Hu, Song, & Allen, 2006).

TFIDF and its variants were applied to the clinical texts such as EMRs (Hoogendoorn, Szolovits, Moons, & Numans, 2016; Napolitano, Marshall, Hamilton, & Gavin, 2016). Semantic features such as named entities and semantic predications were additionally considered exploiting the clinical texts (Kavuluru, Rios, & Lu, 2015). The medical ontology UMLS was employed to identify concepts for the semantic features in EMRs. As clinical texts contain narratives about the patient diagnosis and treatment, the importance of terms in clinical texts closely relates to the patients' status.

While the term importance in a given document identified by frequency-based weighting methods such as TFIDF is fixed, it is worth noting that the semantic importance of that term can be varied depending on the aspect under consideration, in other words the semantic importance is aspect-sensitive. For example, a word in a clinical note (a symptom) can be very important for diagnosing the disease but can be less important in the treatment of the disease.

Beside the variation of the semantic term importance in clinical texts, another challenge is the number of medical terms too large to assign each term a different semantic weight. For example, UMLS (Bodenreider, 2004) comprises over one million biomedical concepts and five million concept names. Distinctly identifying different aspect-sensitive semantic weights to such a huge number of medical terms is infeasible.

Our key idea in addressing those challenges is to divide terms in clinical texts into different categories (nodes) in a hierarchy where terms in the leaf nodes roughly have similar importance, which means that it is assigned with the same semantic weights. By exploiting the essence of hierarchical structure, the deeper the hierarchy, the higher the semantic weights. Those categories will be organized in a two-part hierarchy. The first part consists of categories at high levels of the hierarchy that can be commonly used in different applications. The second part consists of categories flexibly organized by specific medical domain knowledge regarding the aspect under consideration in each application.

To this end, the purpose of this paper is to propose a two-phase framework which generates the two-part hierarchy for determining semantic weights of terms in clinical texts and develop a method for semantic term weighting in EMRs clinical texts regarding the severity of patients' conditions. For the first phase, the categories of terms in the first part of the hierarchy are formed using the medical ontology UMLS as well as ICD-10 codes. For the second phase, we employ a ranking of causes of death (Murphy, Xu, & Kochanek, 2013) which is compatible with ICD-10 to form the second part of the hierarchy as well as the subcategories of the first part regarding the severity of patients' conditions. The semantic weights of the terms in the leaf nodes are assigned in a manner to preserve a decreasing order in the hierarchy and adjusted by parameter Δ. The final weight of a term in a clinical text will be combined with the TFIDF weight and the semantic weight in conjunction with a parameter α.

2. Methods

Our solution for term weighting when considering the semantics of terms is a combination of TFIDF and the semantic weight. Given a term t_i in the document d, the TFIDF weight w_f is computed as follows

$$w_f = TFIDF(t_i, d) = TF(t_i, d) \times IDF(t_i, d) = \frac{n_i}{\sum_k n_k} \times \frac{|D|}{|\{d : t_i \in d\}|} \quad (1)$$

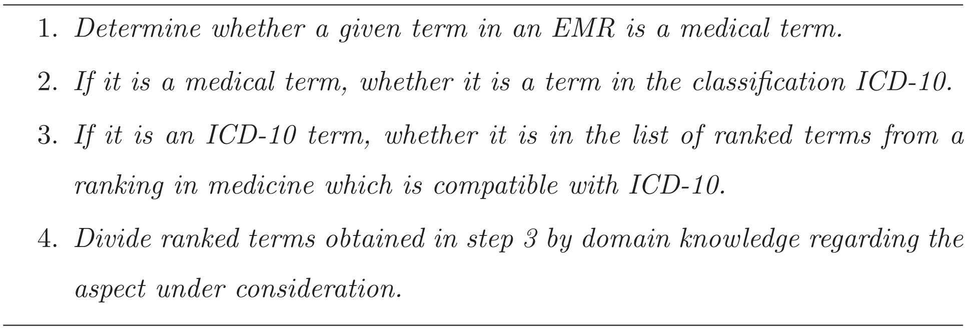
where n_i is the frequency of the term t_i in the document d, $\Sigma_k n_k$ is the sum of the frequency of all terms appearing in the document d, $|D|$ is the total number of documents and $|\{d: t_i \in d\}|$ is the number of documents containing t_i. The TFIDF weight of a term is a number between 0 and 1. If the term appears more frequently in the document and simultaneously appears less frequently in other documents, the TFIDF weight is high. It indicates the term is more important for the document. Given a term in a clinical text, denote by w_f the weight obtained by TFIDF, and denote by w_m the weight obtained by medical importance of the term, the final weight w of the term is defined as

$$w = (1 - \alpha) \times w_f + \alpha \times w_m \quad (2)$$

where $0 \leq \alpha \leq 1$ is a parameter to balance the two weights. This work focuses on computing the medical importance and the effect of the parameter α on the combination of the two weights.

2.1. The framework

The semantic term weighting aims to give a weight to each term in a clinical text according to its medical importance. To this end, our key idea is to employ existing rankings of medical concepts that have been widely used in medicine. In fact, we employ UMLS and ICD-10 for forming the categories of terms having increasing medical importance in the first part of the hierarchy



1. *Determine whether a given term in an EMR is a medical term.*
2. *If it is a medical term, whether it is a term in the classification ICD-10.*
3. *If it is an ICD-10 term, whether it is in the list of ranked terms from a ranking in medicine which is compatible with ICD-10.*
4. *Divide ranked terms obtained in step 3 by domain knowledge regarding the aspect under consideration.*

Fig. 1. Four steps of the two-phase framework.

and use special domain knowledge to refine the category with the highest weight in the second part of the hierarchy. Fig. 1 presents our proposed two-phase framework. The first phase consists of steps 1–3 and the second phase mentioned in step 4 will be described in another section.

In the first step the task of determining whether a given term in an EMR is a medical term is carried out by employing the Unified Medical Language System (UMLS). UMLS is composed of the three main parts of a metathesaurus that are a repository of more than five million of biomedical concepts and their synonyms, a semantic network which provides 135 categories of the concepts as well as lexical resources, and tools for using UMLS resources (Bodenreider, 2004). We firstly use the tool MetaMap of UMLS to map the biomedical text to the UMLS metathesaurus (Aronson, 2001). We then consider the term as a medical term if it has a Concept Unique Identifiers (CUI) code in UMLS and go to the second step, otherwise the term is regarded as a non-medical term and we put it in category C_1. We consider the terms in C_1 do not have any medical importance and assign them the value zero as semantic weights.

In the second step the task is to determine whether the medical term identified in the first step is a term in the classification ICD-10. The International Statistical Classification of Diseases and Related Health Problems (ICD) is an international standard diagnostic classification for all general epidemiological and many health management purposes (World Health Organization, 2004). The classification provides alphanumeric codes of medical terms for diagnoses where the codes are structured in a hierarchy. We utilize CUI codes to obtain ICD-10 codes from the identified medical terms in the previous step. We identify whether the medical term has an ICD-10 code by using the interoperable code of the UMLS concept on BioPortal (Noy et al., 2009), as it can map the CUI code to the ICD-10 classification. BioPortal is an open repository of biomedical ontologies that range in subject matter such as anatomy, phenotype, experimental conditions, imaging, chemistry, and health. BioPortal also represents mappings between terms in different ontologies (Noy et al., 2009). If the term is an ICD-10 term we go to the third step, otherwise it is put into category C_2. We consider the terms in C_2 have some medical importance, but correspond to a low weight as the category contains only general medical terms which are not related to any concrete diseases.

The third step is to determine whether the ICD-10 term identified in the second step is in the list of ranked terms of a certain ranking in medicine which is compatible with ICD-10. The combination of the ranking and the ICD-10 hierarchical structure is accomplished by connecting the ICD-10 code of the ICD-10 term in the hierarchical structure with the ICD-10 code of each rank in the ranking. Thus, the ranking gives the medical importance weights of each rank to the corresponding ICD-10 terms in the hierarchical structure of ICD-10. If the term is not a ranked term, it is put into category C_3. If the term is a ranked term, it is put into category C_R. The terms in C_3 have medical importance higher than the terms in C_2. The categories C_1, C_2 and C_3 can be commonly used for different applications and they form the first part of the hierarchy. The category C_R will then be divided into subcategories based on domain knowledge regarding the aspect under consideration. The next subsection describes the division of C_R into subcategories regarding the disease severity. Assuming C_R will be divided into K-3 categories in the second part of the hierarchy, then the two-part hierarchy has totally K categories that contain the categories C_1, C_2 and C_3 in the first part and the categories C_4, C_5, ... , C_K in the second part.

2.2. Forming subcategories of ICD-10 ranked terms regarding the disease severity

To assess the two-phase framework presented in the previous section, we illustrate the second phase of dividing C_R into subcategories regarding the aspect of the disease severity. To this end, we adopt the widely accepted medical knowledge about the ranking of death causes (Murphy et al., 2013). The statistical information is compiled in a national database through the Vital Statistics Cooperative Program of the Centers for Disease Control and Preventions National Center for Health Statistics (Murphy et al., 2013). The causes of death study ranks diseases into 15 categories with increasing severity relating the patient death. The ICD-10 ranked terms are thus divided into 15 categories in terms of corresponding diseases. Totally, terms in clinical texts are basically divided into 18 categories regarding the disease severity. In the next subsection, we present how the weights are assigned to those 18 categories.

Fig. 2 shows the two-part hierarchy in case of disease severity study consisting of 18 categories where the first part consists of C_1, C_2 and C_3 and the second part consists of C_4, C_5, ... , C_{18}.

Fig. 3 presents some sentences in an EMR clinical text and Fig. 4 describes the classification process of the terms in those sentences into medical importance categories after removal of stop-words. For instance, terms ‘female’ and ‘episode’ belong to category C_1 as non-medical terms according to MetaMap. In contrast, terms ‘paroxysmal nocturnal dyspnea’ and ‘hypercholesterolemia’ belong to category C_2 because these do not have ICD-10 codes even they are medical terms. The term ‘shortness of breath’ is an ICD-10 term which is not ranked in the ranking of causes of death. Accordingly, this term is classified into category C_3. The ICD-10 ranked terms in the death causes ranking will belong to the categories between C_4 and C_{18}. The term ‘congestive heart failure’ where the ICD-10 code is I50 corresponds to the top rank. Hence, this term belongs to category C_{18}. The term ‘diabetes mellitus’ is an ICD-10 ranked term where the ICD-10 code is E10-E14.9. As the term is positioned as the rank 7, it belongs to category C_{12}. The term ‘hypertension’ where the ICD-10 code is I10-I15.9 corresponds to both category C_{18} (rank 1) and category C_6 (rank 13). Thus, there are

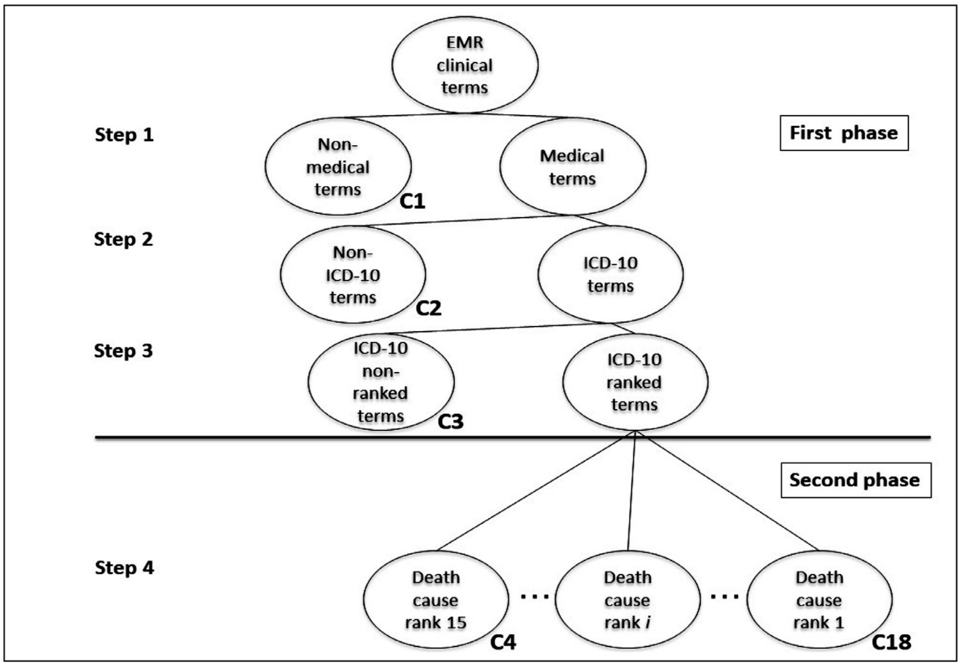


Fig. 2. The illustration to classify each EMR's term into 18 medical importance categories.

1. *Mrs. [**Known patient lastname 4483**] is an 81 year old female with congestive heart failure. She has been medically managed but has gradually experienced worsening symptoms of dyspnea on exertion and paroxysmal nocturnal dyspnea.*
2. *She did have that one episode of shortness of breath which was most likely due to acute pulmonary edema.*
3. *As the patient has risk factors of diabetes mellitus, hypertension, and hypercholesterolemia and possible old inferior myocardial infarction on electrocardiogram it was felt that ischemia was the likely cause of her conduction system abnormalities.*

Fig. 3. Example of sentences in EMRs.

ICD-10 ranked terms which are not classified into a specific category.

2.3. Determination of the semantic weights for each category

The ultimate problem is to appropriately determine the semantic weights for K categories of EMR's terms regarding their medical importance (18 categories in case of disease severity study). Denote by $w(C_i)$ the weight (a real number) to be assigned to category C_i regarding the medical importance of C_i. The essence of term weighting in the proposed method is the increasing order of $w(C_i)$ in the two-part hierarchy but not their absolute values. The determination of $w(C_i)$ should obey the following constraint

Proposition 1. *The values of $w(C_i)$ can be arbitrarily determined but have to preserve the ordinal relation*

$$w(C_1) < w(C_2) < \ldots < w(C_K)$$

From the Proposition where preserving the ordinal relation is essential, we can consider the difference of weights of two consecutive categories as a constant Δ. Since category C_1 does not contain any medical terms, the weight $w(C_1)$ initially is zero. Thus, Δ should satisfy $\Delta \leq \frac{1}{K}$ to ensure $w(C_K) \leq 1$, and the weight of categories is consecutively updated as follows

$$w(C_{i+1}) = w(C_i) + \Delta \quad (3)$$

The value of $w(C_i)$ does not reflect the real importance of terms in the category but it preserves the order of $w(C_i)$. In this work, we consider four degrees: 0.04, 0.03, 0.02 and 0.01 for the parameter Δ. Table 1 indicates the weights of each category with the corresponding name of cause of death as well as the ICD-10 code(s) and the rank. Different category weights can be described by varying the parameter Δ. The average weight of the categories are computed if an ICD-10 ranked term corresponds to multiple ranks.

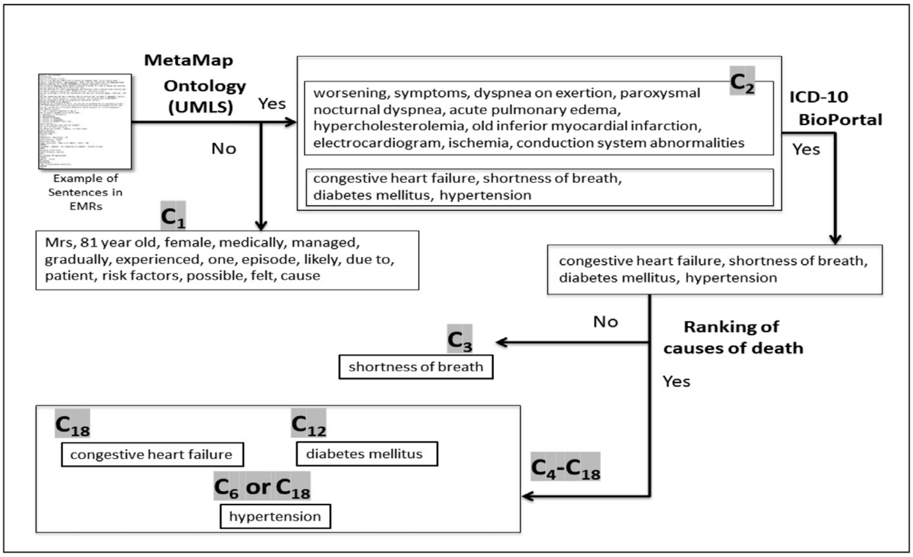

MetaMap
Ontology
(UMLS)
Yes
No
Example of Sentences in EMRs
C2
worsening, symptoms, dyspnea on exertion, paroxysmal nocturnal dyspnea, acute pulmonary edema, hypercholesterolemia, old inferior myocardial infarction, electrocardiogram, ischemia, conduction system abnormalities
congestive heart failure, shortness of breath, diabetes mellitus, hypertension
ICD-10
BioPortal
Yes
C1
Mrs, 81 year old, female, medically, managed, gradually, experienced, one, episode, likely, due to, patient, risk factors, possible, felt, cause
congestive heart failure, shortness of breath, diabetes mellitus, hypertension
No
Ranking of causes of death
Yes
C3
shortness of breath
C18
congestive heart failure
C12
diabetes mellitus
C6 or C18
hypertension
C4-C18

Fig. 4. Example of the classification process of terms appearing in EMRs into medical importance categories.

Table 1
The ranking-based medical importance weights in terms of the severity of patients' conditions.

Rank	Name of cause of death	ICD-10 code(s)	Weight ($\Delta = 0.04$)	Weight ($\Delta = 0.03$)	Weight ($\Delta = 0.02$)	Weight ($\Delta = 0.01$)	Category
1	Disease of heart	I00-I09, I11, I13, I20-I51	0.7	0.7	0.7	0.7	C_{18}
2	Malignant neoplasms	C00-C97	0.66	0.67	0.68	0.69	C_{17}
3	Chronic lower respiratory diseases	J40-J47	0.62	0.64	0.66	0.68	C_{16}
4	Cerebrovascular diseases	I60-I69	0.58	0.61	0.64	0.67	C_{15}
5	Accidents (unintentional injuries)	V01-X59, Y85-Y86	0.54	0.58	0.62	0.66	C_{14}
6	Alzheimer's disease	G30	0.49	0.55	0.6	0.65	C_{13}
7	Diabetes mellitus	E10-E14	0.45	0.52	0.58	0.64	C_{12}
8	Nephritis, nephritic syndrome and nephrosis	N00-N07, N17-N19, N25-N27	0.41	0.49	0.56	0.63	C_{11}
9	Influenza and pneumonia	J09-J18	0.37	0.46	0.54	0.62	C_{10}
10	Intentional self-harm (suicide)	U03, X60-X84, Y87.0	0.33	0.43	0.52	0.61	C_9
11	Septicemia	A40-A41	0.29	0.4	0.5	0.6	C_8
12	Chronic liver disease and cirrhosis	K70, K73-K74	0.25	0.37	0.48	0.59	C_7
13	Essential hypertension and hypertensive renal disease	I10, I12, I15	0.21	0.34	0.46	0.58	C_6
14	Parkinson's disease	G20-G21	0.16	0.31	0.44	0.57	C_5
15	Pneumonitis due to solids and liquids	J69	0.12	0.28	0.42	0.56	C_4
16	ICD-10 non-ranked terms	None	0.08	0.25	0.4	0.55	C_3
17	Medical terms (No ICD-10 code)	None	0.04	0.22	0.38	0.54	C_2
18	Non-medical terms	None	0	0	0	0	C_1

2.4. Combining medical importance weights with TFIDF weights.

The medical importance weight w_m is finally combined with the TFIDF weight w_f for the final weight w under consideration by following Eq. (4)

$$w = (1 - \alpha) \times w_f + \alpha \times w_m \tag{4}$$

where α is a coefficient of the two weights in [0, 1] to adjust the TFIDF weight and the medical importance weight for concrete applications.

Note that the combination of the two weights is executed for terms appearing in clinical texts after preprocessing of the clinical texts such as stop word removal, chunking and removing the terms correspond to negation words.

3. Experimental evaluation

3.1. Objective

This section compares the proposed semantic term weighting method with the TFIDF-based method as a baseline to verify the effectiveness of the proposed framework. Moreover, this section elucidates adequate parameters of Δ and α for the proposed weighting.

3.2. Experimental design

This section conducts an experiment on mortality prediction for the evaluation as the proposed weighting is based on the severity of patients' conditions or death causes. EMRs of elderly patients are used from the well-known database MIMIC II (Saeed et al., 2011). The patients belong to two categories, one is people who died in hospital and the other is people who remained in hospital.

Regarding the statistical aspect, this work uses a total of 13,026 EMRs that contain information about patients who are more than 60 years old. The numbers of EMRs corresponding to the two categories' labels are 2158 and 10,868, respectively. Table 2 describes the distribution of the document frequencies of terms and the number of terms where each term belongs to one of the 18 categories.

We evaluate the proposed method by four options corresponding to four different values of the parameter Δ, namely, TFIDF + MED ($\Delta = 0.04$), TFIDF + MED ($\Delta = 0.03$), TFIDF + MED ($\Delta = 0.02$) and TFIDF + MED ($\Delta = 0.01$). These options are compared to

Table 2
Distribution of the document frequencies of terms and the number of terms in each category.

Category	Sum of document frequencies of terms	Average of document frequencies of terms	Percentage of document frequencies of terms	Number of terms
C_1	4,222,157	324.133	0.840461229	102,068
C_2	675,397	51.8499	0.134444312	12,418
C_3	74,904	5.7503	0.014910366	1507
C_4	0	0	0	0
C_5	1238	0.095	0.000246436	13
C_6	10,304	0.791	0.002051111	24
C_7	389	0.03	7.74342E-05	17
C_8	1340	0.103	0.00026674	19
C_9	106	0.01	2.11003E-05	1
C_{10}	2633	0.202	0.000524124	39
C_{11}	2571	0.197	0.000511782	20
C_{12}	4699	0.361	0.000935381	37
C_{13}	211	0.02	4.20016E-05	12
C_{14}	0	0	0	0
C_{15}	3053	0.234	0.000607729	30
C_{16}	3317	0.255	0.000660281	23
C_{17}	780	0.06	0.000155267	2
C_{18}	20,520	1.5753	0.004084705	239

Table 3
Results using AdaBoost.

α	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
1	0.746	0.745	0.729	0.733
0.9	0.845	**0.854 ($p < 0.01$)**	**0.851 ($p < 0.05$)**	**0.85 ($p < 0.05$)**
0.8	**0.851 ($p < 0.01$)**	0.841	0.848	0.845
0.7	0.842	0.848	0.842	0.844
0.6	0.843	0.843	0.836	0.841
0.5	0.839	0.842	0.843	0.843
0.4	0.835	0.844	0.839	0.843
0.3	0.839	0.844	0.837	0.84
0.2	0.841	0.838	0.835	0.839
0.1	0.829	0.833	0.83	0.838
0	*0.82*	*0.815*	*0.823*	*0.816*

Table 4
Results using Decision Tree.

α	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
1	0.634	0.625	0.619	0.624
0.9	0.785	**0.778**	**0.778 ($p < 0.01$)**	**0.776**
0.8	0.779	0.766	0.769	**0.776**
0.7	0.782	0.771	0.757	0.772
0.6	0.779	0.766	0.772	0.77
0.5	0.78	0.771	0.771	0.769
0.4	**0.787 ($p < 0.1$)**	0.771	0.769	0.764
0.3	0.78	0.773	0.767	0.773
0.2	0.782	0.773	0.767	0.771
0.1	0.775	0.77	0.77	0.768
0	*0.768*	*0.756*	*0.759*	*0.759*

Table 5
Results using Gradient Boosting.

α	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
1	0.776	0.77	0.765	0.763
0.9	0.873	**0.873 ($p < 0.1$)**	0.868	0.869
0.8	**0.875 ($p < 0.01$)**	0.872	0.867	0.867
0.7	0.874	0.871	0.868	0.869
0.6	0.873	0.871	0.869	**0.871 ($p < 0.05$)**
0.5	0.872	0.872	0.868	0.87
0.4	0.874	**0.873 ($p < 0.05$)**	0.868	0.87
0.3	**0.875 ($p < 0.01$)**	0.872	**0.871 ($p < 0.1$)**	0.869
0.2	0.873	**0.873 ($p < 0.05$)**	**0.871**	0.87
0.1	0.874	**0.873 ($p < 0.05$)**	0.868	**0.871**
0	*0.865*	*0.859*	*0.862*	*0.863*

Table 6
Results using Linear Discriminant Analysis.

α	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
1	0.697	0.677	0.676	0.666
0.9	0.776	0.755	0.764	0.751
0.8	0.784	0.78	0.782	0.779
0.7	0.782	0.792	0.799	0.788
0.6	0.784	**0.793 ($p < 0.01$)**	**0.8 ($p < 0.01$)**	0.788
0.5	**0.785 ($p < 0.01$)**	0.789	0.797	0.788
0.4	**0.785 ($p < 0.01$)**	0.792	0.797	0.793
0.3	0.778	**0.793 ($p < 0.01$)**	0.798	0.792
0.2	0.781	0.79	0.798	0.791
0.1	0.76	0.783	0.794	**0.794 ($p < 0.01$)**
0	*0.727*	*0.713*	*0.729*	*0.717*

Table 7
Results using Logistic Regression.

α	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
1	0.694	0.727	0.73	0.722
0.9	0.69	0.728	0.73	0.727
0.8	0.687	0.726	0.726	0.726
0.7	0.685	0.724	0.728	0.729
0.6	0.684	0.72	0.729	0.728
0.5	0.69	0.72	0.728	0.729
0.4	0.697	0.715	0.733	0.726
0.3	0.719	0.715	0.726	0.727
0.2	0.726	0.727	0.723	0.726
0.1	0.714	**0.756 ($p < 0.05$)**	**0.735**	0.724
0	**0.735**	*0.728*	*0.726*	**0.738**

Table 8
Results using Neural Network.

α	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
1	0.672	0.682	0.693	0.69
0.9	0.69	0.691	0.697	0.699
0.8	0.73	0.694	0.706	0.698
0.7	0.75	0.717	0.722	0.699
0.6	0.767	0.734	0.74	0.726
0.5	0.784	0.746	0.751	0.742
0.4	0.803	0.764	0.765	0.755
0.3	0.814	0.787	0.775	0.765
0.2	0.824	0.807	0.8	0.776
0.1	**0.826 ($p < 0.01$)**	**0.824 ($p < 0.05$)**	0.814	0.807
0	*0.818*	*0.808*	**0.817**	**0.82**

Table 9
Results using Naive Bayes.

α	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
1	0.641	0.674	0.684	0.677
0.9	0.647	0.679	0.684	0.678
0.8	0.669	0.691	0.684	0.68
0.7	0.678	0.696	0.687	0.682
0.6	0.69	0.704	0.699	0.684
0.5	0.708	0.709	0.704	0.687
0.4	0.722	0.718	0.718	0.695
0.3	0.746	0.73	0.721	0.703
0.2	**0.754 (p <0.01)**	0.75	0.734	0.719
0.1	0.746	**0.769 (p <0.01)**	**0.76 (p <0.01)**	**0.743 (p <0.1)**
0	*0.731*	*0.726*	*0.717*	*0.728*

Table 10
Results using Random Forest.

α	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
1	0.619	0.606	0.579	0.598
0.9	**0.862 (p <0.01)**	0.841	0.836	0.824
0.8	0.843	**0.846 (p <0.01)**	0.835	0.841
0.7	0.841	0.842	0.837	**0.853 (p <0.01)**
0.6	0.85	0.841	**0.846 (p <0.01)**	0.837
0.5	0.847	0.839	0.829	0.834
0.4	0.838	0.842	0.832	0.845
0.3	0.852	0.838	0.84	0.833
0.2	0.845	0.84	0.835	0.847
0.1	0.833	0.841	0.833	0.843
0	*0.82*	*0.817*	*0.814*	*0.813*

Table 11
Results using SVM(linear).

α	TFIDF + MED ($\Delta = 0.04$)	TFIDF + MED ($\Delta = 0.03$)	TFIDF + MED ($\Delta = 0.02$)	TFIDF + MED ($\Delta = 0.01$)
1	0.744	0.717	0.704	0.685
0.9	0.744	0.723	0.712	0.693
0.8	0.753	0.73	0.722	0.698
0.7	0.761	0.741	0.728	0.708
0.6	0.772	0.752	0.742	0.721
0.5	0.784	0.766	0.755	0.735
0.4	0.792	0.779	0.776	0.752
0.3	0.809	0.79	0.786	0.767
0.2	**0.824 (p <0.01)**	0.805	0.798	0.789
0.1	0.815	**0.819 (p <0.05)**	**0.81 (p <0.05)**	**0.804**
0	*0.795*	*0.792*	*0.789*	*0.798*

Table 12
Comparative results of nine classifiers between the proposed method and the baseline (TFIDF).

Classifier	Baseline	Proposed method
AdaBoost	0.815	**0.854**
Decision Tree	0.768	**0.787**
Gradient Boosting	0.865	**0.875**
Linear Discriminant Analysis	0.729	**0.8**
Logistic Regression	0.728	**0.756**
Neural Network	0.818	**0.826**
Naive Bayes	0.726	**0.769**
Random Forest	0.82	**0.862**
SVM(linear)	0.795	**0.824**
Average	0.785	**0.817**

the baseline (TFIDF) when varying the parameter α. Note that the baseline's result corresponds to the results of the proposed method in its four options where the parameter α is zero.

In this experiment we use nine classifiers: AdaBoost, Decision Tree, Gradient Boosting, Linear Discriminant Analysis, Logistic Regression, Neural Network, Naive Bayes, Random Forest and SVM(linear). Each classifier is executed after the feature selection by using L2 regularization where the parameters of each classification method and L2 regularization are default according to the methods provided by the Scikit-learn toolkit (Pedregosa et al., 2011). We use five trials with a 70% train data and a 30% test data randomly selected from the dataset. Each trial, F1 scores are computed by Eq. (5). To adjust the unbalanced dataset, Synthetic Minority Over-sampling Technique (SMOTE) is employed for the train data with the default parameters of imbalanced-learn (Lemaître, Nogueira, & Aridas, 2017). Therefore, the ratio of the train data is equalized between the two categories' labels.

$$F1\ score = \frac{2 \times precision \times recall}{precision + recall} \tag{5}$$

A Python module Scikit-learn (Pedregosa et al., 2011) is employed in these experiments. A paired *t*-test is carried out by using Scipy (Jones, Oliphant, Peterson et al., 2001), to assert the superiority of the proposed method.

3.3. Experimental results

The result of the baseline (TFIDF) is compared to the result of the proposed method's options: TFIDF + MED ($\Delta = 0.04$), TFIDF + MED ($\Delta = 0.03$), TFIDF + MED ($\Delta = 0.02$) and TFIDF + MED ($\Delta = 0.01$) when varying the parameter α. The results are in Tables 3–12. The *p*-values ($<$ 0.1, $<$ 0.05 and $<$ 0.01) between the highest F1 score of each proposed method's option and the F1 score of the baseline are indicated in the tables.

Overall, using the proposed semantic term weighting with TFIDF can have higher F1 score than only using TFIDF with its significant difference derived from a paired *t*-test. The highest F1 score in those experiments was 87.5% derived from TFIDF + MED ($\Delta = 0.04$) where Gradient Boosting was employed as the classifier and α was 0.8. There was a great difference of the scores between the proposed semantic term weighting and the baseline. For example, the score's difference was approximately 8% where LDA was employed as the classifier. The baseline was partially better than the proposed semantic term weighting when Logistic Regression and Neural Network were used as the classifier.

4. Discussion

The experimental results showed that the proposed semantic term weighting method when varying the parameters was better than the TFIDF-based method. This suggests that the proposed method of semantic term weighting based on the severity of patients' conditions is appropriate for the mortality prediction task.

On the whole, the higher Δ is, the greater the prediction performance. The higher α, the higher the F1 score where AdaBoost and Random Forest were employed as the classifier. On the other hand, the smaller α, the higher the F1 score where Logistic Regression, Neural Network, Naive Bayes and SVM(linear) were employed as the classifier. In those experiments, ensemble learning such as Gradient Boosting, AdaBoost and Random Forest outperformed other classifiers. As for each of the nine classifiers, we compared the highest score of the proposed method with the score of the baseline where the corresponding Δ of that proposed method was employed. The result showed that the proposed method improved approximately 3% of the average F1 score among the nine classifiers. This suggests that the proposed method does not depend on classifiers. Since the default parameters of the classifiers were used according to the methods provided by the Scikit-learn toolkit (Pedregosa et al., 2011), it is also assumed that the proposed method does not depend on the parameters of classifiers.

As the proposed semantic term weighting method was developed based on the medical ontology UMLS, the medical classification ICD-10 and the ranking of causes of death, the proposed method's results were based solely on medical knowledge.

Although methods for mortality prediction have been developed by using scores such as sequential organ failure assessment (SOFA) and simplified acute physiology score (SAPS) or some algorithms without the scores (Houthooft et al., 2015; Jiménez, Sánchez, & Juárez, 2014; Richards, Rayward-Smith, Sönksen, Carey, & Weng, 2001; Ripoll, Vellido, Romero, & Ruiz-Rodríguez, 2014), methods of document representation, the so-called term weighting where clinical data are represented in a vector space by the terms' weights was not exploited for mortality prediction.

It must be considered that the results of the proposed semantic term weighting is strongly effected by the dataset of MIMIC II used in the experiment, as the proposed semantic term weighting gave semantic weights to terms appeared in the dataset. One limitation of the proposed method is based on the performance of MetaMap which was exploited to identify whether a term is medical term or not. The other limitation is the coverage of the ranking which was exploited to identify whether a term is an ICD-10 ranked term or not.

5. Conclusions

In this paper, we proposed a two-phase framework which derives a two-part hierarchy for determining semantic weights of terms in clinical texts and developed a semantic term weighting method for EMRs clinical texts regarding the severity of patients' conditions. The first phase aims to classify all terms into common categories at high levels of the two-part hierarchy by using the medical ontology in UMLS as well as ICD-10. The second phase aims to flexibly classify terms into categories based on the first part of the hierarchy, organized by specific medical domain knowledge regarding the aspect under consideration. We employed a ranking of causes of death to form the subcategories of ICD-10 ranked terms' category in the first part regarding the severity of patients' conditions. The semantic weights of the terms in the leaf nodes of the two-part hierarchy were assigned in a manner to preserve a decreasing order in the hierarchy. The difference of the semantic weights was adjusted by parameter Δ. The final weight of a term in a clinical text was combined with the TFIDF weight and the semantic weight in conjunction with the parameter α.

The proposed framework was evaluated with an implementation for study of severity of patients' conditions where a ranking of death causes was used to identify categories in the second part of the hierarchy. The experimental results of mortality prediction using nine classifiers showed that the proposed method in varying the parameters outperformed the TFIDF-based method. Its effectiveness was verified by a paired *t*-test because there was a significant difference between the proposed method and the TFIDF-based method in terms of their performance. In comparison with the highest score of the proposed method and the scores of the baseline where the corresponding Δ of that proposed method was employed, the proposed method improved approximately 3% of the average F1 score among the nine classifiers.

The proposed two-phase framework can be applied to different tasks in medical domain, by extending the second part of the hierarchy when employing appropriate medical knowledge for the tasks under consideration. The proposed semantic term weighting method can be applied to various prediction tasks regarding patients' risk or severity-based similar case retrieval on clinical texts such as EMRs, because the proposed method exploited the ranking of causes of death which contains 15 ranks regarding diseases.

Our proposed approach for semantic term weighting simply represents clinical texts into the vector space model form (like document-term matrix) by transforming a term into a semantic weight regarding the aspect under consideration. In the process of the representation, the proposed approach can generate a two-part hierarchy as a knowledge base that organizes a huge amount of distinct terms in clinical texts with its semantic weights regarding the aspect under consideration in the categories of the hierarchy. Therefore, in the prevalence of EMRs, the proposed approach contributes to pervasive the exploitation of clinical texts such as EMRs in various applications regarding medicine and also share and integrate clinical data in different systems for healthcare management.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments. This work is partially supported by Vietnam National University at Ho Chi Minh City under the grant number B2016-42-01.

Declarations of interest: None.

References

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: The metamap program.. In *Proceedings of the AMIA symposium* (pp. 17–21). American Medical Informatics Association.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research, 32*(suppl 1), D267–D270.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199–220.

Hoogendoorn, M., Szolovits, P., Moons, L. M., & Numans, M. E. (2016). Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. *Artificial Intelligence in Medicine, 69*, 53–61.

Houthooft, R., Ruyssinck, J., van der Herten, J., Stijven, S., Couckuyt, I., Gadeyne, B., et al. (2015). Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores. *Artificial Intelligence in Medicine, 63*(3), 191–207.

Jiménez, F., Sánchez, G., & Juárez, J. M. (2014). Multi-objective evolutionary algorithms for fuzzy classification in survival prediction. *Artificial Intelligence in Medicine, 60*(3), 197–219.

Jing, L., Zhou, L., Ng, M. K., & Huang, J. Z. (2006). Ontology-based distance measure for text clustering. In *Proceedings of SIAM SDM workshop on text mining, Bethesda, Maryland, USA*.

Jones, E., Oliphant, T., Peterson, P. et al. (2001). SciPy: Open source scientific tools for Python. [Online; accessed July 11, 2018] http://www.scipy.org/.

Kavuluru, R., Rios, A., & Lu, Y. (2015). An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial Intelligence in Medicine, 65*(2), 155–166.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research, 18*(17), 1–5.

Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems with Applications, 38*(10), 12708–12716.

Murphy, S. L., Xu, J., & Kochanek, K. D. (2013). Deaths: Final data for 2010. *National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System, 61*(4), 1–117.

Napolitano, G., Marshall, A., Hamilton, P., & Gavin, A. T. (2016). Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. *Artificial Intelligence in Medicine, 70*, 77–83.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., et al. (2009). Bioportal: Ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research, 37*, W170–W173.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(Oct), 2825–2830.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. *Technical report*. Department of Computer Science, Rutgers University.

Richards, G., Rayward-Smith, V. J., Sönksen, P., Carey, S., & Weng, C. (2001). Data mining for indicators of early mortality in a database of clinical records. *Artificial Intelligence in Medicine, 22*(3), 215–231.

Richesson, R. L., Sun, J., Pathak, J., Kho, A. N., & Denny, J. C. (2016). Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artificial Intelligence in Medicine, 71*, 57–61.

Ripoll, V. J. R., Vellido, A., Romero, E., & Ruiz-Rodríguez, J. C. (2014). Sepsis mortality prediction with the quotient basis kernel. *Artificial Intelligence in Medicine, 61*(1), 45–52.

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., et al. (2011). Multiparameter intelligent monitoring in intensive care ii (mimic-ii): A public-access intensive care unit database. *Critical Care Medicine, 39*(5), 952–960.

Sakre, M. M., Kouta, M. M., & Allam, A. M. (2009). Weighting query terms using wordnet ontology. *International Journal of Computer Science and Network Security, 9*, 349–358.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523.

Sureka, V., & Punitha, S. (2012). Approaches to ontology based algorithms for clustering text documents. *International Journal of Computer Technology and Applications, 3*(5), 1813–1817.

Tar, H. H., & Nyunt, T. T. S. (2011). Ontology-based concept weighting for text documents. *World Academy of Science, Engineering and Technology, 81*, 249–253.

Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E. G., & Milios, E. E. (2005). Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on web information and data management* (pp. 10–16). ACM.

World Health Organization (2004). *International statistical classification of diseases and related health problems*: 1. World Health Organization.

Yang, C. C., & Veltri, P. (2015). Intelligent healthcare informatics in big data era. *Artificial Intelligence in Medicine, 65*(2), 75–77.

Yu, H., & Cao, Y.-G. (2009). Using the weighted keyword models to improve information retrieval for answering biomedical questions. *AMIA summit on translational bioinformatics*.

Zakos, J., & Verma, B. (2006). Concept-based term weighting for web information retrieval. *International Journal of Computational Intelligence and Applications, 6*(02), 193–207.

Zhang, X., Jing, L., Hu, X., Ng, M., Jiangxi, J. X., & Zhou, X. (2008). Medical document clustering using ontology-based term similarity measures. *International Journal of Data Warehousing and Mining (IJDWM), 4*(1), 62–73.

Zhang, X., Jing, L., Hu, X., Ng, M., & Zhou, X. (2007). A comparative study of ontology based term similarity measures on pubmed document clustering. In *International conference on database systems for advanced applications* (pp. 115–126). Springer.

Zhu, W., Xu, X., Hu, X., Song, I.-Y., & Allen, R. B. (2006). Using umls-based re-weighting terms as a query expansion strategy. In *Ieee international conference on granular computing* (pp. 217–222).