Artificial Intelligence in Medicine 51 (2011) 27–41

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Conceptual-driven classification for coding advise in health insurance reimbursement

Sheng-Tun Li [a,*], Chih-Chuan Chen [c], Fernando Huang [b]

[a] Institute of Information Management, National Cheng Kung University, Department of Industrial and Information Management, National Cheng Kung University, No. 1, Ta-Hsueh Road, Tainan 701, Taiwan, ROC
[b] Institute of Information Management, National Cheng Kung University, No. 1, Ta-Hsueh Road, Tainan 701, Taiwan, ROC
[c] Department of Leisure Information Management, Taiwan Shoufu University, No.168, Nanshih Li, Madou, Tainan 721, Taiwan, ROC

ARTICLE INFO

Article history:
Received 27 March 2009
Received in revised form 16 June 2010
Accepted 8 October 2010

Keywords:
Knowledge discovery
Text mining
Fuzzy formal concept analysis
Information retrieval
ICD code
Health insurance

ABSTRACT

Objective: With the non-stop increases in medical treatment fees, the economic survival of a hospital in Taiwan relies on the reimbursements received from the Bureau of National Health Insurance, which in turn depend on the accuracy and completeness of the content of the discharge summaries as well as the correctness of their International Classification of Diseases (ICD) codes. The purpose of this research is to enforce the entire disease classification framework by supporting disease classification specialists in the coding process.
Methodology: This study developed an ICD code advisory system (ICD-AS) that performed knowledge discovery from discharge summaries and suggested ICD codes. Natural language processing and information retrieval techniques based on Zipf's Law were applied to process the content of discharge summaries, and fuzzy formal concept analysis was used to analyze and represent the relationships between the medical terms identified by MeSH. In addition, a certainty factor used as reference during the coding process was calculated to account for uncertainty and strengthen the credibility of the outcome.
Results: Two sets of 360 and 2579 textual discharge summaries of patients suffering from cerebrovascular disease was processed to build up ICD-AS and to evaluate the prediction performance. A number of experiments were conducted to investigate the impact of system parameters on accuracy and compare the proposed model to traditional classification techniques including linear-kernel support vector machines. The comparison results showed that the proposed system achieves the better overall performance in terms of several measures. In addition, some useful implication rules were obtained, which improve comprehension of the field of cerebrovascular disease and give insights to the relationships between relevant medical terms.
Conclusion: Our system contributes valuable guidance to disease classification specialists in the process of coding discharge summaries, which consequently brings benefits in aspects of patient, hospital, and healthcare system.

1. Introduction

Taiwan's National Health Insurance program was formally launched in March 1995 and has since provided medical care and treatment to the entire population of Taiwan under the principle of mutual assistance and risk sharing (see http://www.nhi.gov.tw/, accessed 27 March 2009). The whole community pays insurance premiums that constitute the income of the Bureau of National Health Insurance (BNHI) and is ensured of receiving financial support on a wide range of medical services, which explains the ideal of outstanding cheap healthcare for all.

Following the policies established by the BNHI, the original Retrospective Payment System (RPS) provided unlimited reimbursement to medical service providers in Taiwan according to the cost associated with each service provided. However, this induced service providers to overuse medical resources to increase their income and in addition, the aging of the Taiwanese population required improvements in medical care quality and increases in the amount of drugs covered by the payment system. This caused the mean growth rate in revenue of recent years to be less than the growth rate in medical costs and lead to a financial crisis. In order to reduce the deficit, the RPS, which operated on a fee-for-service basis, was replaced with a Global Budget Payment System

* Corresponding author. Tel.: +886 6 2757575x53126; fax: +886 6 2362162.
E-mail addresses: stli@mail.ncku.edu.tw (S.-T. Li), r3895101@mail.ncku.edu.tw (C.-C. Chen), fernandohuang@gmail.com (F. Huang).

0933-3657/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2010.10.003

(GBPS) that operates on a case payment basis. The GBPS reimburses all medical service providers from a limited global budget consistent with the amount of cases handled instead of the cost incurred, reducing overuse and waste. The reimbursement system under GBPS is based on the Diagnosis Related Group (DRG), a system that categorizes hospital cases into groups according to the International Classification of Diseases (ICD) and assigns weights to each group that indicate the amount of resources necessary to treat a patient with a given diagnosis.

In most Taiwanese hospitals, discharge summaries are analyzed by disease classification specialists and assigned codes according to the patients' diagnoses in order to classify them into one of the many existent DRGs, each associated with a fixed mean patient cost. Reimbursements are made to service providers on a per-discharge basis the value originally appointed, disregarding the resources used in each case, so that service providers control their usage. The coding of discharge summaries is a complex task that requires careful analysis and revision, so is usually done by senior disease classification specialists. However, when workloads are heavy or the coding is being done by novice personnel, it might become a tedious and error-prone task. Any mistake in the coding process may lead to the misclassification of discharge summaries and cause incorrect reimbursement, so medical service providers have become extremely severe with the coding process so as to ensure that discharge summaries are correctly coded and to avoid any economic loss. Consequently, it is essential to guarantee the quality of disease classification specialists and reduce the probability of error.

The purpose of this research is to enforce the entire disease classification framework by supporting disease classification specialists in the coding process. This decreases the economic loss of hospitals and facilitates the management and distribution of funds by the BNHI. In order to accomplish this, an ICD code advisory system (ICD-AS) that performs conceptual clustering of discharge summaries and recommendation of ICD codes is developed. ICD-AS also reveals inherent relationships between medical concepts through knowledge representation and implication rules, improving the correctness and quality of the coding process. We utilize fuzzy formal concept analysis integrated with natural language processing and feature extraction techniques to elicit medical knowledge embedded in discharge summaries of patients suffering from cerebrovascular disease.

The structure of the paper is organized as follows. Section 2 reviews underlying technologies on formal concept analysis and information retrieval techniques and Section 3 describes the specifics of the research methodology. Section 4 illustrates the system development process and Section 5 presents the results of the experiments. Section 6 discusses performance evaluation and finally, Section 7 draws conclusions on the paper and provides insights on future work.

2. Underlying technologies

2.1. Formal concept analysis

Formal concept analysis (FCA) is a mathematical approach proposed by Wille [1] used for structuring, analyzing and visualizing data, based on a notion of duality called Galois connection [2]. Data consisting of a set of entities and their features are structured into formal abstractions called formal concepts, which together form a concept lattice ordered by a partial order relation. Concept lattices are constructed by identifying the objects and their corresponding attributes for a specific domain, called conceptual structures, and then displaying the relationships between them.

In FCA, a context is represented as a triple (O, A, R) where O is a set of objects, A is a set of elements and R is a binary relation between O and A. If for an object o and an attribute a, oRa holds, then we say that "object o has attribute a", or "attribute a applies to object o". Table 1 shows an example of a context for hospital discharge summaries and their respective topics (represented by attributes in FCA), where each row is an object representing a discharge summary, each column is a topic, and each X represents the existence of the relationship.

Table 1
Context for hospital discharge summaries.

	Cancer	Fever	Cold	Leukemia	Sore throat
DS 1				X	
DS 2		X	X		X
DS 3	X	X		X	
DS 4		X			
DS 5				X	
DS 6		X			
DS 7			X		X
DS 8			X		X
DS 9			X		X
DS 10		X	X		X

Two sets can be derived from a given formal context (O, A, R), one from the perspective of objects and another from the viewpoint of attributes. The former is a subset of objects $I \subseteq O$ such that I' is defined by the attributes that apply to all objects belonging to I and the latter is a subset of attributes $J \subseteq A$ such that J' is defined by the objects that have all attributes belonging to J. Correspondingly,

$$I' = \{a \in A | oRa \quad \forall o \in I\},$$

$$J' = \{o \in O | oRa \quad \forall a \in J\},$$

thus the pair (I, J), where $I \subseteq O$ and $J \subseteq A$, is a formal concept of the context (O, A, R) if $I' = J$ and $J' = I$. The formal concept (I, J) is a pair where I consists of those objects that have all attributes from the set J, and equally, J consists of those attributes that apply to all objects from the set I. Additionally, the set of objects is known as the extent of the concept and the set of attributes as the intent.

Taking the context for hospital discharge summaries as an example, from the point of view of objects, objects DS 2 and DS 10 have the same attributes "fever", "cold" and "sore throat"; on the other hand, from the attributes' perspective, attributes "fever", "cold" and "sore throat" apply to objects DS 2 and DS 10. Thus {{DS 2, DS 10}, {"fever", "cold", "sore throat"}} is a formal concept in which the extent is the set of objects {DS 2, DS 10} and the intent is the set of attributes {"fever", "cold", "sore throat"}. The formal concept is derived by identifying the relevant objects and their relevant features, thus it is a clustering of objects and their related common attributes obtained by observation of the reality [3].

Given a formal context (O, A, R), an inheritance relation $(\leq)$ between two concepts $\{I1, J1\}$ and $\{I2, J2\}$ can be established according to the following condition:

$$\{I_1, J_1\} \leq \{I_2, J_2\} \quad \text{iff} \quad I_1 \subseteq I_2, J_2 \subseteq J_1.$$

There exists a natural hierarchical order between the concepts of a given context, in this case $\{I1, J1\}$ is a subconcept of $\{I2, J2\}$ and $\{I2, J2\}$ is a superconcept of $\{I1, J1\}$. An object i will have an attribute j if and only if there is an upwards leading path from the concept having i to the concept having j [4].

The set of all concepts in context (O, A, R) and their inheritance relations form the concept lattice, in which nodes are labeled with the concepts of the context and arcs are used to join the associated nodes that are in $\leq$ relation. The $\leq$ relation is a partial order relation that expresses a double inclusion among node components. Increasing the number of objects in a formal concept will reduce the number of components in the intent, implying a specialization process. Conversely, increasing the number of attributes in a for-

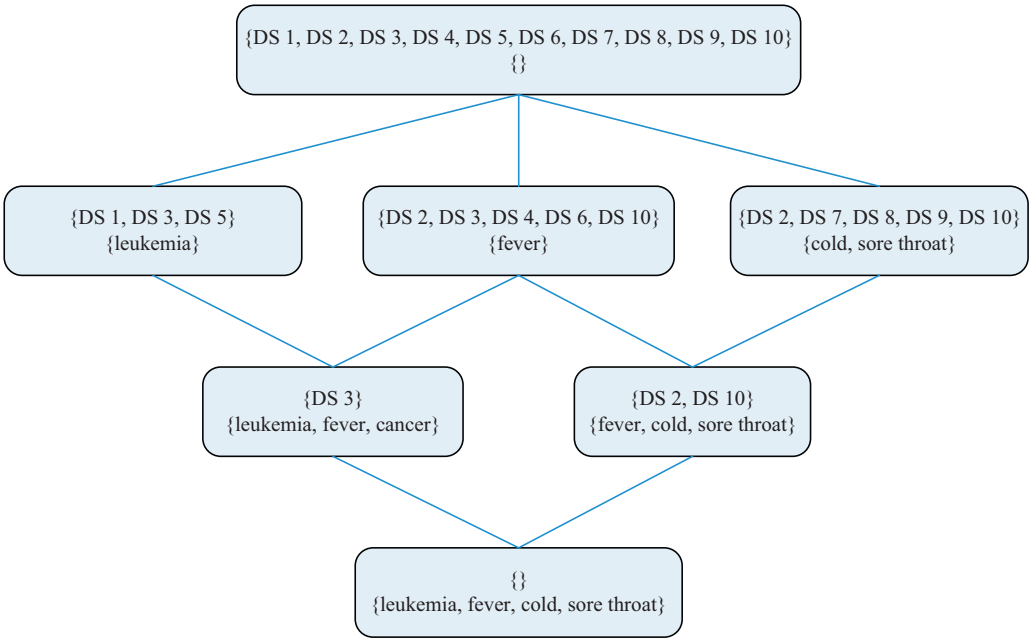

Fig. 1. Concept lattice of context for hospital discharge summaries.

mal concept will reduce the number of components in the extent, implying a generalization process.

In addition, a concept lattice has two special nodes that set the boundaries of the entire lattice, specifically the maximum and minimum nodes, which respectively represent the most general and most specific concepts of the lattice. For each node, its extent is contained in all of its ancestors and its intent contains the intent of all of its ancestors. In other words, the attributes are inherited from the most general maximum node, while the objects are inherited from the most specific minimum node. Consequently, a concept lattice will contain redundant information, but is very useful because both parts of the concept can be alternatively used in different situations. Objects sharing the same attributes will be found in the same formal concept and those which only share some or more attributes, in neighboring linked formal concepts. This coincides with the idea of clustering based on feature similarity and allows grouping of documents into a lattice to facilitate search and retrieval [5]. Fig. 1 shows the concept lattice obtained from the context for hospital discharge summaries (Table 1).

Furthermore, a concept lattice contains a set of implications between attributes that can be elicited in the form of rules. An implication rule between two attribute sets (expressed as $J1 ===> J2$, where J_1, $J_2 \subseteq A$ and $J_1 \cap J_2 = \varnothing$) represents a relation between them such that any object set having attribute set $J1$ also has attribute set $J2$. Hence any hidden domain knowledge in the concept lattice can be revealed through the analysis of implication rules.

FCA has been used in information retrieval, knowledge discovery in databases, knowledge representation [2]. Related works on FCA include using implication rules calculated by FCA to support case-based reasoning applications on a travel agency [6]. Jiang et al. [7] explored the role of FCA in a context-based ontology building support in a clinical domain and proved the correctness of the implication rules obtained, while Boucher-Ryan and Bridge [8] applied FCA to identify the relation between users and items and find neighbors in a collaborative recommender system. In addition, Carpineto and Romano [9] supported information retrieval applications with FCA to perform query refinement, create conceptual representations of documents and represent retrieval results in browsable concept lattices. Tho et al. [10] built relations between clusters and represented them using FCA in order to improve citation-based document retrieval systems into finding research domain expertise, and Rajapakse and Denham [11] investigated the use of more realistic concepts for document retrieval and reinforcement learning for improving document representation to help the retrieval of useful documents for relevant queries based on FCA. Formica [3] proposed the evaluation of similarity between concepts of a given context using FCA and ontologies and Weng et al. [12] also stressed the importance of FCA in the automatic generation of ontological concepts to facilitate the exchange, search and identification of text information. Tho et al. [13] combined fuzzy theory with FCA to extend its capability of manipulating uncertain data and performed conceptual clustering to construct a concept hierarchy and support document retrieval. Fuzzy theory was also implemented by Belohlavek et al. [14] to achieve factorization of concept lattices to reduce computation time and reduce lattice complexity.

There are also some works applying FCA on ICD-related studies. Jiang et al. proposed a model for formalizing ICD coding rules underlying the ICD Index using FCA. The coding rules were generated from FCA models and represented in the Semantic Web Rule Language (SWRL) [15]. Jay et al. applied FCA to explore cancer patient flows in the French region of Lorraine with an easily understandable visual metaphor [16]. Data were extracted from the year 2003 DRG database of the Lorraine region in which hospital stays related to cancer care through the use of selected codes from the ICD were identified. They found that FCA, as an unsupervised conceptual clustering method, could describe patients flows with an easily understandable visual representation than the traditional approach based on geographical information systems. Upon reviewing the literature regarding FCA applications, we have not seen studies concerning insurance reimbursement.

2.2. Fuzzy formal concept analysis

Fuzzy formal concept analysis (FFCA) was introduced by Tho et al. [13] by incorporating fuzzy logic into FCA in order to represent vague information. Traditional FCA uses a binary approach indicating the presence or absence of the relation between an object and an attribute and cannot take uncertain information into consideration, while FFCA extends the relation into a value in the interval

Table 2
Fuzzy formal context for hospital discharge summaries.

	Cancer	Fever	Cold	Leukemia	Sore throat
DS 1	0.16	0.10	0.00	0.98	0.00
DS 2	0.02	0.81	0.76	0.13	0.79
DS 3	0.87	0.75	0.00	0.83	0.07
DS 4	0.13	0.89	0.14	0.35	0.09
DS 5	0.32	0.02	0.00	0.92	0.00
DS 6	0.08	0.67	0.07	0.22	0.33
DS 7	0.00	0.01	0.71	0.00	0.78
DS 8	0.00	0.07	0.59	0.00	0.89
DS 9	0.00	0.19	0.85	0.00	0.60
DS 10	0.10	0.62	0.90	0.30	0.98

of real numbers between 0 and 1 called membership value. In this way, the relation is not simply belongs or does not belong anymore, but can take a set of values indicating a degree of belonging. This membership value of an attribute towards an object is proportional to its significance and can be used to calculate a similarity value between two formal concepts. These notions are represented in the following definitions [13,17].

Definition 1 *(Fuzzy formal context).* A fuzzy formal context is a triple $(G, M, F = \varphi(G \times M))$, where G is a set of objects, M is a set of attributes, and F is a fuzzy relation between G and M. Each relation $(g, m) \in F$ has a membership value $\mu(g, m)$ in [0,1].

Definition 2 *(Fuzzy formal concept).* A fuzzy formal concept of a fuzzy formal context (G, M, F) is a pair $(\varphi(Y), Z)$ for $Y \subseteq G$ and $Z \subseteq M$, where each object $g \in \varphi(Y)$ has a membership value $\mu_Y(g) = \min_{m \in Z} \mu(g, m)$ with $\mu(g, m)$ being the membership value of $(g, m) \in F$.

Definition 3 *(Fuzzy formal concept cardinality).* The fuzziness of a fuzzy formal concept is represented by membership values of objects of the concept, hence the cardinality of a fuzzy formal concept $(\varphi(Y), Z)$ is defined as $|\varphi(Y)|$, where $|\varphi(Y)|$ is the cardinality of the fuzzy set $\varphi(Y)$.

Definition 4 *(Fuzzy formal concept similarity).* The similarity of a fuzzy formal concept $(\varphi(Y_1), Z_1)$ and its subconcept $(\varphi(Y_2), Z_2)$ is defined as

$$E((\varphi(Y_1), Z_1), (\varphi(Y_2), Z_2)) = E(\varphi(Y_1), \varphi(Y_2)) = \frac{|\varphi(Y_1) \cap \varphi(Y_2)|}{|\varphi(Y_1) \cup \varphi(Y_2)|},$$

where $\cap$ and $\cup$ are the intersection and union operations of fuzzy sets $\varphi(Y_1)$ and $\varphi(Y_2)$, respectively.

There is a wide variety of ways of defining intersection and union operations of fuzzy sets. In this paper, we choose commonly used ways. That is, $\varphi(Y_1) \cap \varphi(Y_2)$ is defined as $\mu_{Y_1 \cap Y_2}(g) = \min(\mu_{Y_1}(g), \mu_{Y_2}(g))$, and $\varphi(Y_1) \cup \varphi(Y_2)$ is defined as $\mu_{Y_1 \cup Y_2}(g) = \max(\mu_{Y_1}(g), \mu_{Y_2}(g))$.

Table 2 is a fuzzy formal context represented as a cross-table, which shows the relations between a set of objects representing ten discharge summaries and a set of attributes representing topics "cancer", "fever", "cold", "leukemia" and "sore throat". It also contains membership values that indicate the significance of the topic for each discharge summary, in contrast to the formal context in Table 1 which only shows the existence of the relation.

The similarity between two linked fuzzy formal concepts, namely C1 and C2 in a fuzzy concept lattice can be calculated using the following formula:

$$E(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}, \quad (1)$$

as defined in Definitions 3 and 4. Formica [3] proposed another formula to calculate concept similarity, but it is not suitable to this research as it requires a similarity graph between attributes that has to be previously established by a domain expert.

2.3. Information retrieval

The field of information retrieval focuses mainly on the search for documents in large databases and retrieval of unstructured data in response to a query. Due to the overwhelming amount of unstructured data and growing number of textual information available, the need for effective information retrieval methods, especially from text documents, has made this discipline a very popular center of attention [18].

2.3.1. Zipf's law

Zipf's law is a scientific law based on mathematical statistics that has been applied in information retrieval for feature selection purposes. The basic concept is that the frequency of any word is inversely proportional to its rank in the frequency table, the frequency of the ith most frequent word will be as many times as that of the most frequent one divided by i^r, for some $r > 1$, producing a long tail distribution [19]. Luhn investigated the impact of frequency of words on the resolving power in discriminating a given text or document and found that the middle frequent words were the significant words since high and low frequency words are generally very ambiguous or not important enough in the text, respectively [20]. Luhn described the relationship between the resolving power of terms and Zipf's law as shown in Fig. 2 and suggested that less discriminating words could be removed by establishing upper and lower frequency cutoffs [20].

The terms, ranked according to their frequency, are divided into three groups delimited by an upper and lower cutoff.

- *Stop words group:* Terms belonging to this group are located to the left of the upper cutoff, have the highest frequencies and show a very low resolving power. The likelihood that two terms in this group have similar frequencies is very low, in fact, their frequen-

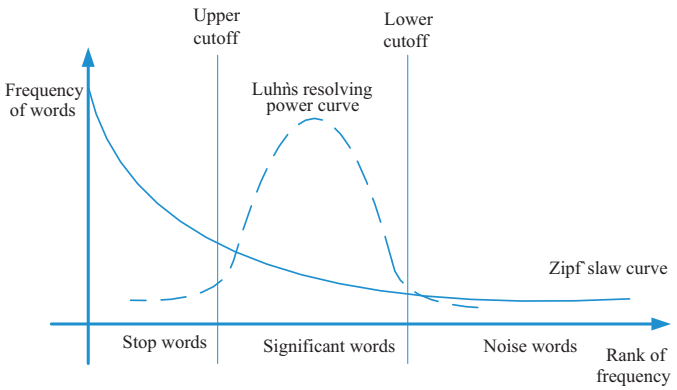

Fig. 2. Luhn's resolving power and Zipf's law.

cies vary greatly. The frequencies of terms in this group can be linearly regressed with slope approximating negative infinity.
- *Noise words group:* Terms belonging to this group are found to the right of the lower cutoff and present low frequency as well as low resolving power. The likelihood that two noise words have the same frequencies is very high, producing a long tail distribution. The frequencies of terms in this group can be linearly regressed with slope approximating zero.
- *Significant words group:* This group consists of the words surrounded by the upper and lower cutoffs, they present the highest resolving power and are the most significant and representative terms of a document or corpus. The frequencies of terms in this group can be linearly regressed with slope approximating a value between zero and negative infinity.

Zipf's law has been widely applied among different fields and has proven to possess many uses, such as explaining the occurrence of words in texts, proving distributions of populations in cities, etc. With the growth of the Internet, the connections of networks, connectivity of routers, web caching strategies many other elements of the WWW model have proven to follow a Zipf distribution [21]. An analysis of long term queries to websites on Russian academic networks showed that the rank distribution of websites was approximated by Zipf distribution, suggesting that website popularity is a universal property of the Internet and follows Zipf's law [22]. Furthermore, the distribution of low variability periods in the activity of human heart rate has also proven to follow Zipf's law [23].

2.3.2. Term weighting

Terms describe the content of their text sources to a certain extent, which is reflected on a numeric weight value that represents the relative significance of the term and acts as an estimate of the usefulness of the given term as a descriptor of the given document. Numerous term weighting functions have been proposed and tested [24], which most depend on statistical methods or on the distribution of the terms in the documents.

The most successful and widely used scheme for automatic generation of weights is the "term frequency inverse document frequency" (TF-IDF) weighting function. A term may be a better descriptor of one document than of another, hence it may have different weight values for each document in which it occurs. When it occurs frequently in the entire corpus it is a poor indicator of its source, such as frequent stop-words or the term animal in a collection of documents about animals. But when it appears frequently in few documents, it is a discriminative term and will have a higher TF-IDF value. In other words, the fewer the term appears on other texts the more discriminative it is, so its weight will be inversely related to the number of documents in which it appears. In this way, whenever a term appears in many different documents, its weight will decrease together with its relative significance. Thus, the terms that best describe a document are those that appear frequently in it but not in others. The TF, or term frequency of a term, represents the frequency of occurrence of the term in a specific document. For term i in a specific document,

$$tf_i = \text{frequency of occurrence of index term } i \text{ in the text.} \quad (2)$$

The IDF, or inverse document frequency of a term, represents the significance of the term in the entire corpus. It is obtained by dividing the total amount of documents in the corpus by the number of documents containing term i and then taking the logarithm. For term i in a specific document,

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}, \quad (3)$$

where D is the total number of documents in the corpus and $\{d : t_i \in d\}$ the number of documents containing term i. Then we have

$$TF\text{-}IDF = tf \times idf. \quad (4)$$

As a result, a term with high frequency in a document and low frequency in other documents of the corpus will yield a higher TF-IDF value, indicating its high significance in the document.

3. Research method

This research uses discharge summaries of hospital patients as corpus, which are preprocessed by natural language processing and filtered by the Medical Subject Headings (MeSH) database and clinical experts to identify relevant medical terms. Subsequently, feature extraction is performed based on Zipf's law to identify those medical terms which are significant in the corpus and fuzzy formal concept analysis (FFCA) is employed to analyze the relationships between the final selection of relevant medical terms. The lattice diagram produced is used to recommend ICD codes for discharge summaries and the implication rules are used to understand the inherent relationships between the medical terms. An overview of the proposed ICD code advisory system, which is basically a medical decision support system, is shown in Fig. 3. It follows the framework of medical DSS building blocks suggested by [25]. Note that the components of Fuzzy FCA and ICD code suggestion can be replaced with other traditional classification methods such as decision tree, Naïve Bayes, multilayer perceptrons, support vector machine (SVM), etc.

3.1. Natural language processing

Documents containing data in natural language format are useful sources of information, but due to the several variations in human language, it is problematical for automated systems to accurately understand its meaning. Natural language processing (NLP) consists of any attempt to process natural language using computers, but has faced numerous problems and limitations since in order to really understand its connotation, extensive knowledge about the outside world is required. The general process starts by separating textual input into tokens, followed by stemming to map related words to the same root and reduce the amount of tokens, stop-word elimination, part-of-speech tagging and finally syntax analysis [26].

This research identifies medical terms from discharge summaries that will later be used as attributes that describe each discharge summary. The process consists of two main steps: *tokenization* and *stemming*. The first step performs splitting of textual content into tokens by eliminating all punctuation marks and blank spaces to obtain words, while the second step carries out a reduction of words into their stem forms using the MeSH database. Common stop words such as "the", "for", "a", "an", etc. are also simply discarded in the process.

3.2. Medical term generation

MeSH is a large vocabulary system created and maintained by the United States National Library of Medicine (NLM) with the aim of assisting cataloging, indexing, and searching for biomedical documents (http://www.nlm.nih.gov/mesh/, accessed 27 March 2009). Each main heading, also known as descriptor, is selected by clinical experts according to their degree of common use and general acceptance. It indicates the subject of an indexed item and contains a short definition, is linked to other related descriptors, and has a list of synonyms called entry terms that allow users to find the most

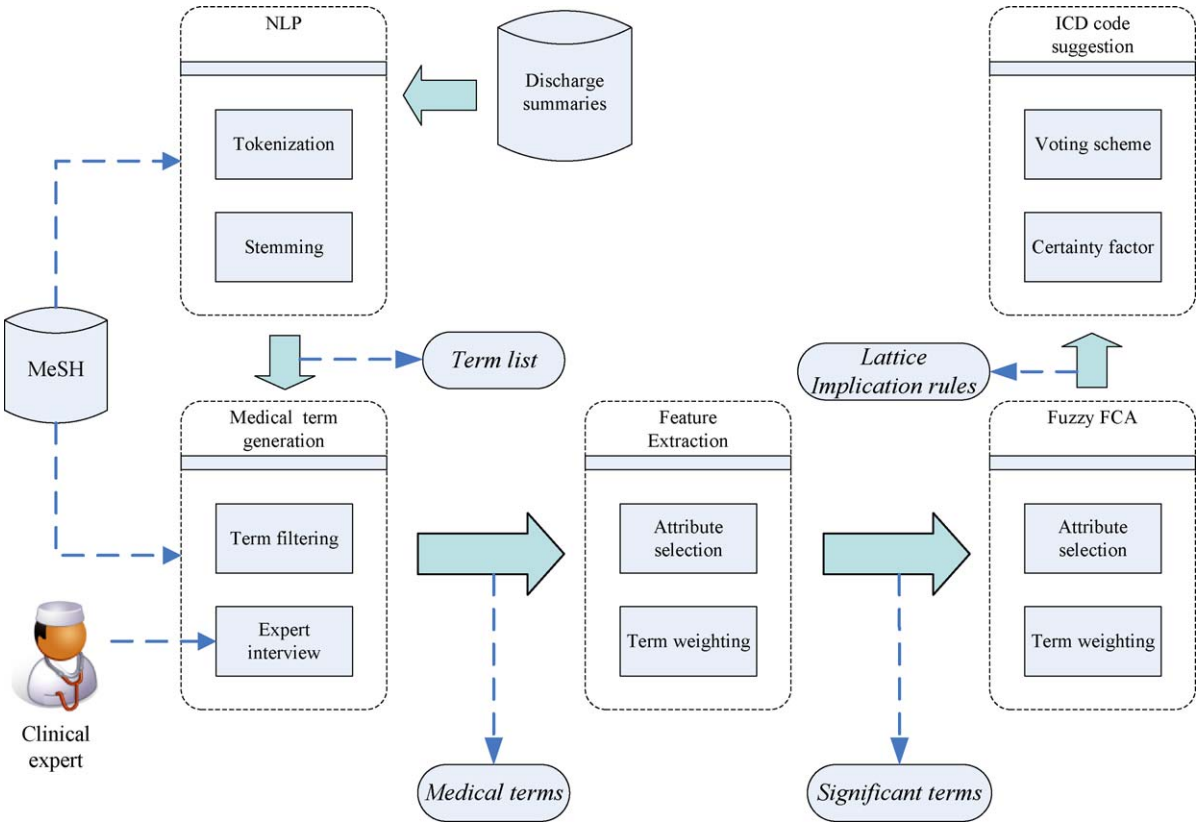

Fig. 3. The proposed ICD-AS.

fitting descriptors for the concepts they are seeking. MeSH also contains a number of qualifiers, or subheadings, which can be used in conjunction with descriptors to narrow down the topic searched. The mapping of all similar terms into the same subject heading facilitates the communication between the different fields inside medical science and the indexing and retrieval of documents. The entire collection is widely used in the biomedical area, including the MEDLINE article database and NLM's catalog of book holdings.

In this research we use MeSH as the main filter for the large amount of words obtained from the tokenization process as [27]. Each word that is found in the database is regarded as a medical term; hence its descriptor is taken into consideration for the next step in the framework leaving out any redundancies and irrelevant terms that could increase computational time. In order to additionally narrow the focus of analysis, clinical expertise is consulted to further reduce the number of terms considered by allowing experts to select those terms which are more representative of the domain or are of special interest. Feature extraction is then performed on the resulting medical terms to identify those which are significant in the corpus and should be included in the lattice.

3.3. Feature extraction

The set of medical terms is composed by root terms of medically relevant terms found in the corpus. However, each presents a different degree of significance and must undergo a feature extraction process that involves two steps. First, attribute selection based on Zipf's law is performed to select from the pool of terms those which are most significant and representative. Second, normalized TF-IDF weight function is applied to identify the degree of significance of each term to each discharge summary. Both steps are explained in the following sub-sections.

3.3.1. Attribute selection

FCA uses as input a set of objects and attributes to perform data analysis and build conceptual lattices. Most recorded applications deal with small numbers of objects and attributes, in which case complexity is low and usage of the data produced like lattice browsing or query expansion is easy. However, when FCA is applied to exploration of data of large sizes, issues of complexity and scalability become crucial. As a result, feature selection plays an important role in our system, selecting from the set of medical terms only those that are significant and eliminating stop words and noise words that are irrelevant and at the same time increase complexity and computation time. Conventional information retrieval systems utilize TF-IDF weighting scheme to perform term selection because it identifies the significant terms for each document. However, this method is not effective in our system because it eliminates the inheritance of terms between discharge summaries that is extremely essential in FCA, hence we opt for another term selection method.

Following the concept of Zipf's law, potential significant keywords can be identified from the entire set of words by first recognizing the upper and lower cutoffs; which can be approximated from the diagram of rank of frequency against frequency (Fig. 2) and the diagram of term frequency against number of terms (Fig. 4); respectively. For long tail distributions it holds that for any function $f(k)$ with long tail characteristics; the variance of $f(k)$ will vary dramatically. That is when k is close to zero it appears that

$$f(k) \approx r \times f(k+1), \tag{5}$$

where r is the ratio $f(k)/f(k+1)$.

Based on this notion, we can approximately identify the upper and lower cutoffs that bound the significant words group. When approaching the upper cutoff, all terms are ranked according to

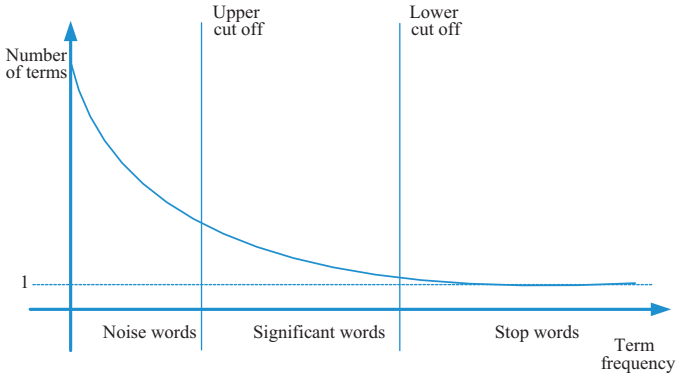

Fig. 4. Term frequency against number of terms.

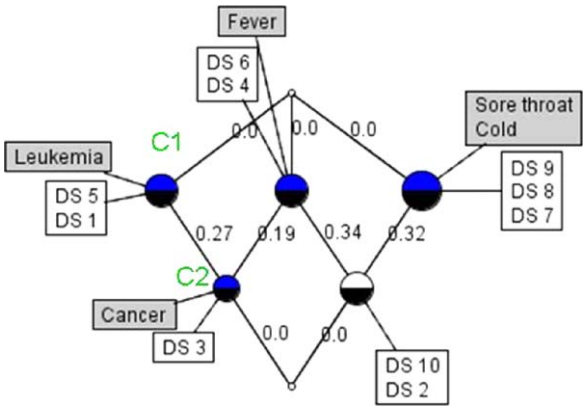

Fig. 5. Example of fuzzy concept lattice.

their frequencies (Fig. 2) and we can identify the upper cutoff by sequentially comparing the frequencies of two adjacent terms starting from the term with the highest frequency. If the ratio $f(k)/f(k+1)$ is smaller than a ratio threshold c set by users, then k is the upper cutoff and all terms having frequency higher than $f(k)$ are considered stop words and discarded.

Similarly, when approaching the lower cutoff, we plot term frequency against number of terms (Fig. 4), and sequentially compare the number of terms of two adjacent term frequencies. If the ratio $f(k)/f(k+1)$ is smaller than a ratio threshold c, then k is the lower cutoff and all terms having frequency lower than k are considered noise words and are discarded.

Once the upper and lower cutoffs are identified and the significant words group is clearly separated from the noise words group and stop words group, we use term weighting function to calculate the accurate degree of significance of each of these significant terms.

3.3.2. Term weighting

To evaluate how important each significant term is to every discharge summary, we apply TF-IDF weight function to compute a value that represents the degree of significance of the term in each discharge summary. However, documents have different sizes and the frequency of a term might vary proportionally to the document's size, in other words, the frequency of a term might be larger for long documents and smaller for short ones and the real significance of the term would be neglected. To overcome this issue, the TF-IDF value of a term can be normalized according to the length of the document by dividing it by the Euclidean length of the document vector. This process is called cosine normalization and is defined by the following equation [28]:

$$\text{Normalized } tf_i\text{-}idf_i = \frac{tf_i\text{-}idf_i}{\sqrt{\sum (tf_j\text{-}idf_j)^2}}, \quad (6)$$

where tf_i-idf_i is the TF-IDF value of term i and tf_j-idf_j is the TF-IDF value of all other distinct index terms in the document.

Once the terms that will comprise the lattice attribute set are identified, their normalized TF-IDF values are calculated to obtain a degree of significance to each discharge summary in the lattice object set. These significant terms are subsequently used as input for fuzzy FCA.

3.4. Fuzzy formal concept analysis

FCA has been applied in various fields like document analysis, ontologies, e-learning, case-based reasoning, etc., but when input data is excessive, the formal concepts formed and the relationships between them are extremely complex and difficult to understand. Moreover, FCA does not provide any insights on the degree of the relationships between attributes and objects, thus fuzzy theory is incorporated.

The normalized TF-IDF values previously calculated for each term represent the degree of significance of the attributes to the discharge summaries and are used as membership values to build up the fuzzy formal context. An α-cut acting as a membership degree threshold is set to eliminate relations that have low membership degree and in this way, reduce accordingly the complexity of the analysis process. All membership values below the α-cut are considered equivalent to zero such that unnecessary noise is removed. Table 3 shows the fuzzy formal context in Table 2 with α-cut = 0.5, where the null values signify the absence of the topic or its exclusion due to low significance. Fig. 5 shows the fuzzy concept lattice built from Table 3, with similarity values between each linked formal concept computed using formula (1) to help identify different concepts that are semantically close. For example, let Y_1 = {DS 1, DS 3, DS 5}, Y_2 = {DS 3}, Z_1 = {"Leukemia"}, Z_2 = {"Fever", "Cancer", "Leukemia"}, then we have $C_1 = \{Y_1, Z_1\}$, $C_2 = \{Y_2, Z_2\}$. The similarity

Table 3
Fuzzy formal context in Table 2 with α-cut = 0.5.

	Cancer	Fever	Cold	Leukemia	Sore throat
DS 1	–	–	–	0.98	–
DS 2	–	0.81	0.76	–	0.79
DS 3	0.87	0.75	–	0.83	–
DS 4	–	0.89	–	–	–
DS 5	–	–	–	0.92	–
DS 6	–	0.67	–	–	–
DS 7	–	–	0.71	–	0.78
DS 8	–	–	0.59	–	0.89
DS 9	–	–	0.85	–	0.60
DS 10	–	0.62	0.90	–	0.98

between fuzzy formal concepts $C1$ and $C2$ is calculated as follows:

$$\begin{aligned} E(C_1, C_2) &= E(\varphi(Y_1), \varphi(Y_2)) \\ &= \left| \frac{\varphi(Y_1) \cap \varphi(Y_2)}{\varphi(Y_1) \cup \varphi(Y_2)} \right| \\ &= \frac{\min[\mu_{Y_1}(DS3), \mu_{Y_2}(DS3)]}{\max[\mu_{Y_1}(DS1), \mu_{Y_2}(DS1)] + \max[\mu_{Y_1}(DS3), \mu_{Y_2}(DS3)] + \max[\mu_{Y_1}(DS5), \mu_{Y_2}(DS5)]} \\ &= \frac{0.75}{0.98 + 0.87 + 0.92} = 0.27 \end{aligned}.$$

The similarity computed between two formal concepts, together with the lattice diagram and the implication rules calculated from the lattice serve as reference for specialists in the disease classification process, contributing to a better understating of the discharge summaries and a more accurate coding.

3.5. Evaluation

FCA provides a means of structuring, analyzing and visualizing data. Formal concepts are composed by those objects that have all attributes of the intent and those attributes that apply to all objects of the extent. In other words, objects sharing the same attributes will be found in the same concept and those partially similar will appear in interrelated neighboring concepts. This notion of clustering based on feature similarity allows documents with similar content to be grouped together and those with distinct content to be located farther. Hence, a document will possess the same features as nearby documents, allowing the possibility to guess or predict its features based on its adjacent neighbors. Analogously, discharge summaries having similar contents will share the same ICD codes and be found in neighboring concepts, while those having distinct contents will be assigned different ICD codes and found in further concepts.

Whenever a discharge summary of interest is present in a formal concept, similar discharge summaries can be traced by following the edges of the concept. Going upwards, or looking into more general superconcepts, will yield discharge summaries containing less attributes, while going downwards, or looking into more specific subconcepts, will yield discharge summaries containing more attributes. Discharge summaries in subconcepts are more specific and are described by more terms, meaning that they contain attributes other than those contained by the discharge summary of interest and are irrelevant. On the other hand, discharge summaries in superconcepts are more general and are described by fewer terms, indicating that these discharge summaries are more "generally similar" to the discharge summary of interest. Discharge summaries with similar contents will be found in close formal concepts, thus the ICD code of a discharge summary can be predicted by observing the codes of those discharge summaries present in the formal concept and its superconcepts. However, discharge summaries of different ICD codes might also share common attributes and a formal concept might contain discharge summaries of diverse ICD codes, hence the process of ICD code prediction is a multiclass problem where a single ICD code must be chosen among many.

This research performs ICD code prediction based on a single-winner voting scheme. For a certain discharge summary in a formal concept, a strength value is calculated for each code present among all "generally similar" discharge summaries in all superconcepts, and the code with the highest strength value is the winner. For each code c, the strength value S is computed as follows:

$$S_c = \sum_{n=1}^{N} \sum_{d_{n,c}=1}^{D} w_{d_{n,c}}, \tag{7}$$

where n is the nth superconcept excluding the root of the lattice, $d_{n,c}$ is a discharge summary of code c occurring in concept n (excluding those discharge summaries also contained by children of concept n that are also superconcepts of the formal concept containing the discharge summary of interest), and $w_{d_{n,c}}$ is the product of all weights on the path between the most specific concept containing $d_{n,c}$ and the concept containing the discharge summary of interest.

Additionally, a certainty factor (cf) is calculated to measure the degree of belief of the winner ICD code. Certainty factors theory was first introduced in MYCIN, an expert system for the diagnosis and therapy of blood infections and meningitis [29], and is used to measure how accurate, truthful or reliable a predicate is. A positive cf represents a degree of belief while a negative cf a degree of disbelief. The maximum possible value is +1.0, meaning definitely true, and the minimum is −1.0, meaning definitely false. Ishibuchi et al. [30] used certainty factors to evaluate the grade of certainty of fuzzy IF-THEN rules in expert systems, the knowledge base consisted of a set of IF-THEN rules with the following syntax:

IF <evidence>
THEN <hypothesis> {cf}

where the cf represents the belief in the hypothesis when the evidence has occurred. For a given classification problem, the hypothesis becomes more certain when patterns of one particular class appear more often than those of other classes. The resulting hypothesis is determined by the class with the largest strength value, so we find the winner class c_{winner} such that

$$S_{c_{winner}} = \max[S_1, S_2, \ldots, S_T], \tag{8}$$

where T is the total number of classes. If a certain class takes the maximum strength value it becomes the winner class and its cf can be computed using the equation below:

$$CF_{c_{winner}} = \frac{S_{c_{winner}} - S_a}{\sum_{c=1}^{T} S_c}, \tag{9}$$

where $S_a = \sum_{c=1, c \neq c_{winner}}^{T} S_c/(T-1)$. If all instances of a dataset belong to only one class, the cf will be large and it is completely certain that a new datum will belong to this class, but when the instances are evenly distributed among all classes, the strengths of all classes will be similar and the cf will be low, meaning that a new datum will be easily misclassified.

In this research, each ICD code represents a class and its strength value is calculated according to the similarity between concepts and the occurrence of the discharge summaries, such that the more discharge summaries of a certain ICD code there are and the closer these discharge summaries are to the formal concept containing the discharge summary of interest, the larger the strength of that ICD code. Furthermore, a cf is obtained to support the degree of belief of the winner ICD code, such that the smaller the value of the cf the lower the degree of belief. This serves as a reference for disease classification specialists when considering the ICD code suggested by the advisory system for a certain discharge summary.

4. System development

The main goal of our system is to build an ICD code advisory system (ICD-AS) that provides knowledge support to disease classification specialists and improves the coding of discharge summaries. The system was built based on Concept Explorer 1.3 [31], an open source software that implements basic functionalities of FCA and is well known due to its simple interface, ease of use and free

Diagnosis at hospitalization

C4-5 subluxation

Diagnosis at discharge

Subluxation and Herniation of interverteberal disc, cervical c4-5

Main complaint

Right upper limb weakness for two months

Medical history

Present medical history

This 57 y/o male patient suffered from right upper limb weakness for two months. The symptoms became more serious in recent weeks. C-spine MRI was performed and C4-5 subluxation was diagnosed. Then he was referred to our OPD .Throughout the whole course of the disease, there was no fever, no paralysis and no urine incontinence. Under the impression of C4-5 subluxation, he was admitted for surgical treatment.

Past medical history

Hypertension: (-)

DM: (-)

Operation history: L3-5 spondylolisthesis post posterior fixation 10 years ago
Right L/3 ureteral stone S/P URS-SM on 930317

Systemic disease: nil

Personal medical history

Allergy: Not known allergic history

Smoking: (+) for 30 years one pack/day

Drinking: (-)

Family medical history

no DM, hypertension nor cancer

Fig. 6. Sample of partial discharge summary.

of charge. Parts of its components were modified to include fuzzy membership values, display of similarity values between formal concepts, and perform ICD code suggestion supported by degree of belief.

The input to the system is a training set of 240 randomly picked discharge summaries of patients from a medical center in Southern Taiwan suffering from cerebrovascular disease. The discharge summaries were verified and approved by the NHI, meaning that their contents were checked by disease classification specialists to confirm the correctness of the ICD codes. This indicates that the contents of the discharge summaries are consistent with the codes, ruling out any incomplete information or coding errors. The distribution of the discharge summaries is even among six different ICD codes, namely 430, 431, 434, 436, 437 and 438. The fact that this research used the same amount of discharge summaries of each code to perform analysis is inconsistent with real world situations, but evidence shows that many learning algorithms suffer reduced performance when learning from imbalanced data. Whenever the amount of instances of a certain class in a dataset is greatly larger than the amount of the other classes in the set, using these data to perform any kind of learning might yield biased results. To tackle this problem, sampling techniques are applied to obtain different data distributions when dealing with imbalanced data, but all aim to obtain equal amounts of instances of each class, such as random oversampling or random undersampling [32,33]. Hence, in order to avoid the negative effects of sampling data, the same amount of discharge summaries per code is used. Fig. 6 shows a sample of a discharge summary with partial omission due to confidentiality issues.

To build the concept lattice, discharge summaries were used as objects and medical terms obtained from the discharge summaries were used as attributes. But scalability and complexity are important limitations in FCA, hence we performed attribute selection based on Zipf's law and term weighting using normalized TF-IDF to optimize performance and reduce computation time. When performing attribute selection, we obtained 2 as the approximate upper cutoff, leaving out the only term with frequency higher than the second most frequent term: "disease". This is quite understandable as "disease" is a very general term in the medical field, just like stop word "the" in an English document. Furthermore, we also obtained 5 as the approximate lower cutoff, leaving out 221 terms having frequencies lower than 5. The normalized TF-IDF was computed for the resulting focal terms and a membership degree threshold was set to eliminate those relations which were too insignificant.

5. Experiment results

240 textual discharge summaries were processed by our system to produce 307 distinct words that were identified as medical terms, of which 277 were identified by the MeSH database as root terms after eliminating any redundant terms and consulting clinical expertise. These non-redundant words that describe the corpus were used as attributes and the discharge summaries of the corpus as objects to produce the lattice diagram partially shown in Fig. 7. The complexity of the lattice diagram makes its navigation and browsing a very difficult task, emphasizing the ability of the proposed system of automatically suggesting ICD codes and revealing hidden knowledge through implication rules.

5.1. Implication rules

The lattice diagram produced graphically shows connections between relevant attributes which contain knowledge that could be

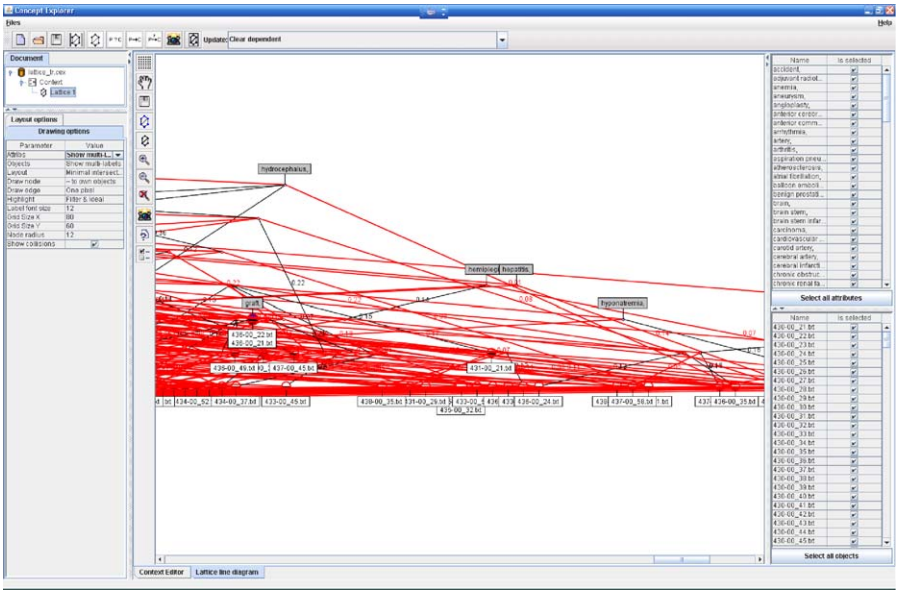

Fig. 7. Partial lattice line diagram built using discharge summaries of patients suffering from cerebrovascular disease.

useful if properly extracted. But these connections are very complex and difficult to understand due to the large amount of attributes and interrelationships present, so association rules were calculated to facilitate knowledge extraction. However, an extremely huge amount of rules were obtained, namely 1259, due to the large number of attributes and complex relations between them, hence a constraint was set to select only those rules with confidence value of 1.0, called implication rules. A rule with a confidence value of 1.0 means that for 100% of the times the antecedent of the rule occurs, its consequent follows and the association is always true. 311 implication rules were obtained and their medical validity was carefully evaluated by a senior expert having 33 years of field experience to ensure the correctness of the knowledge uncovered and identify the type of relationship implied. Relationships are not just empty connectives, but constitute important conceptual units, make significant contributions to meaning, and allow structuring of the entire conceptual framework. Three categories of relationships are widely recognized: relationships of equivalence (ex. definition), of hierarchy (ex. is-a, part-of), and of association (ex. cause–effect), of which are divided into further sub-types [34].

Table 4 lists examples of implication rules calculated the proposed system, which are approved by the expert and which accumulate a 79.17% coverage of the instances. In Table 4, the symbol "===>" represents the implication relation, occurrence stands for the number of instances in which the rule is true, and relationship is the type of relationship the rule corresponds to. For example, the rule "aneurysm, hemorrhage, rupture ===> spontaneous subarachnoid hemorrhage" corresponds to a relationship of association (cause–effect) and is explained by the fact that spontaneous subarachnoid hemorrhage is usually caused by a ruptured aneurysm. However, there are rules that link composite words with their constituent words and do not correspond to any relationship type, hence are labeled as "other".

The lattice line diagram contains knowledge relevant to disease classification hidden in the discharge summaries, which was successfully extracted through implication rules and with the help of clinicians. In this way, disease classification specialists can use this knowledge as reference material for consult during the coding process, especially when checking the content of discharge summaries for correctness or incompleteness.

5.2. ICD code prediction

FCA is a mathematical approach to data analysis based on lattice theory that is able to identify groups of objects with shared properties, which can be further used to cluster similar objects. In our experiment, discharge summaries were chosen at random from the original corpus and an exact copy of each was made, with the file name modified, and put back into the original corpus. The lattice line diagram produced showed that all original discharge summaries and their respective copies were found in the same formal concepts, meaning that discharge summaries having identical content were clustered together. In addition, those with similar but not identical content also appeared in interrelated neighboring formal concepts since they shared common features. Hence, objects with similar attributes were grouped

Table 4
Examples of implication rules calculated by the proposed system.

Rule	Relationship	Occurrence
Urinary tract ===> infection, urinary tract infection	Definition	14
Diabetes mellitus type ii ===> diabetes mellitus	Is-a	45
Spontaneous subarachnoid hemorrhage ===> hemorrhage	Is-a	33
Cerebral infarction ===> infarction	Is-a	17
Coronary artery ===> artery, coronary artery disease	Part-of	19
Cerebral artery ===> artery	Part-of	18
Aneurysm, hemorrhage, rupture ===> spontaneous subarachnoid hemorrhage	Cause–effect	12
Artery, cerebral artery, diabetes mellitus, middle cerebral artery ===> infarction	Cause–effect	7
Artery, hemorrhage, rupture ===> spontaneous subarachnoid hemorrhage	Cause–effect	6
Coronary artery disease ===> artery, coronary artery	Other	19
Total coverage = 79.16%		

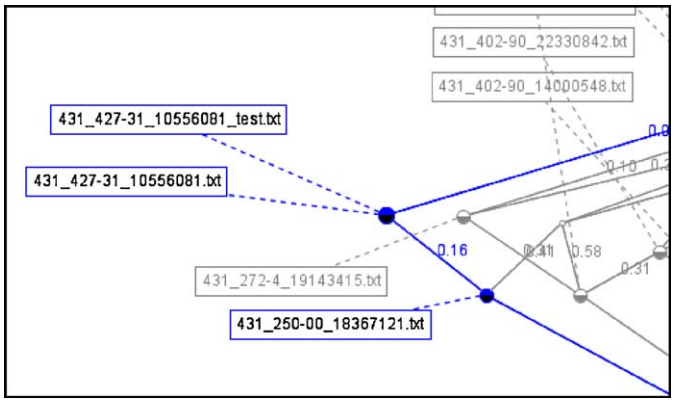

Fig. 8. Identical discharge summaries clustered in the same formal concept.

together in the same formal concepts and those slightly less similar were found in neighboring formal concepts, proving the ability of the advisory system to cluster discharge summaries based on content analysis and perform ICD code prediction. Fig. 8 shows identical discharge summaries clustered in the same formal concept.

By applying fuzzy FCA, similarity values between formal concepts were calculated that provide support and guide disease classification specialists in the coding process. Fig. 9 shows the advisory system suggesting an ICD code for a test discharge summary. Suppose that a discharge summary of code 431-00 named "!!!431-00_TEST" is clustered in the formal concept labeled A and its code wants to be predicted. Notice that formal concepts B, C and R are its superconcepts, but R is excluded because it is the root of the lattice, so we look into the objects of formal concepts A, B and C. Each object in formal concepts A contributes a strength of 1.0 to its respective ICD code, each object in formal concepts B contributes a strength of 0.3, and each object in formal concepts C contributes a strength of 0.38. The winner ICD code is 431-00 with certainty factor 0.47, which is reasonable as Fig. 9 shows that the majority of the discharge summaries considered belong to this ICD code. Hence, the advisory system accurately suggests ICD code 431-00 for the test discharge summary, which coincides with its original code.

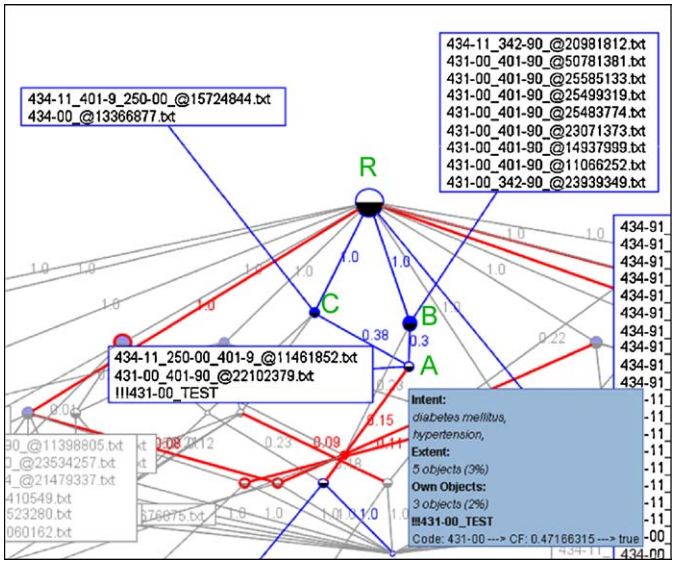

Fig. 9. Advisory system suggesting ICD code for test discharge summary.

6. System evaluation

To evaluate our ICD code advisory system, two sets of discharge summaries are collected. In this section we first evaluate ICD-AS by using the data set afore-discussed and next conduct performance comparison with SVM by using a larger set of discharge summaries.

6.1. Discharge summary set 1

On the basis of constructing ICD-AS as presented in Section 4, we collected another set of 120 discharge summaries, belonging to each of the six ICD codes of the training set, as testing set to test for prediction accuracy. Different experiments were conducted to evaluate the upper and lower cutoffs approximated by Zipf's law, membership degree threshold, and compare the proposed model to traditional classification techniques in terms of prediction accuracy.

6.1.1. Experiment A. Evaluation of membership degree threshold

Following Zipf's law, values of the upper and lower cutoffs were approximated to 2 and 5, respectively, to identify significant words and eliminate stop words and noise words. By setting the upper and lower cutoffs to the estimated values, ICD code prediction was performed using different membership degree thresholds to yield different prediction accuracies. Results show that membership degree threshold and prediction accuracy are indirectly proportional, which is supported by the fact that as the membership degree threshold was increased, more relations were left out making the lattice less specific and less able to perform accurate predictions. Maximum accuracy was achieved for a membership degree threshold of 0.00, which is in this case logical as all noise words and stop words had already been eliminated and there was no need to further exclude any relations.

Table 5 illustrates prediction accuracies using different combinations among upper cutoff, lower cutoff and membership degree threshold, where highlighted values indicate the maximum prediction accuracy. For cutoffs closer to the values approximated following Zipf's law, maximum prediction accuracy was achieved by a membership degree threshold of 0.00. This is consistent with the previously proved fact that membership degree threshold and prediction accuracy are indirectly proportional because noise words and stop words had already been eliminated and no irrelevant relations needed to be eliminated.

It is worth noting that for cutoffs closer to 1, maximum accuracy was achieved by larger membership degree thresholds, which does not coincide with previous facts. However, this discrepancy is correct because for cutoffs closer to 1, more attributes are taken into consideration, which means more stop words and noise words are included in the lattice. Under these circumstances, a membership degree threshold of 0.00 would take all relations into consideration, of which some may be irrelevant and hinder prediction, causing prediction accuracy to drop. A membership degree threshold larger than 0.00 would eliminate insignificant relations caused by the presence of stop words and noise words and would increase prediction accuracy.

6.1.2. Experiment B. Evaluation of upper and lower cutoffs

The upper and lower cutoffs identified using Zipf's law are only approximate values, hence this research experimented using different values for upper cutoff and lower cutoff to find an optimal combination that would yield the best system performance. Membership degree thresholds ranging from 0.00 to 0.50 we tested to find the average prediction accuracy for each combination of upper and lower cutoff (Table 6).

Experimental results show that by setting the upper cutoff to 3, which is close to the value of 2 previously estimated using Zipf's

Table 5
Prediction accuracy for different upper and lower cutoffs and membership degree threshold.

Upper cutoff	Membership degree threshold	Lower cutoff				
		1	2	3	4	5
1	0.00	80.34%	79.66%	81.36%	79.66%	81.36%
	0.10	81.20%	77.97%	78.81%	78.81%	80.51%
	0.15	82.05%	81.51%	78.15%	77.31%	77.50%
	0.20	82.05%	79.66%	77.31%	78.15%	79.17%
	0.25	80.34%	79.49%	75.21%	76.92%	74.79%
	0.30	78.63%	79.49%	77.12%	76.27%	76.47%
2	0.00	76.92%	82.20%	79.66%	79.66%	81.36%
	0.10	78.63%	81.36%	79.66%	80.51%	81.36%
	0.15	81.20%	84.87%	78.99%	78.15%	79.17%
	0.20	78.63%	80.51%	75.63%	77.31%	78.33%
	0.25	78.63%	76.92%	72.65%	74.36%	72.27%
	0.30	76.92%	78.63%	74.58%	75.42%	76.47%

Note: Results for membership degree thresholds larger than 0.30 are not shown, but are consistent with the decreasing trend.

law, prediction accuracy can be maximized. This yields an average accuracy of 75.44% given different lower cutoffs, and a second highest accuracy of 76.77% when the lower cutoff was 4. An upper cutoff of 3 indicates that the two most frequent terms, namely "disease" and "cardiovascular disease", are considered stop words and thus discarded. This is logically correct as the former is a general term in the entire medical field and the latter is a common term in the area of cerebrovascular disease.

Regarding the lower cutoff, evidence shows that by setting its value to 4, close to the value of 5 previously estimated using Zipf's law, yields an average accuracy of 75.42% given different upper cutoffs and a maximum accuracy of 76.77% when the upper cutoff is 3. Although the average accuracies of using lower cutoffs 1 and 2 are only slightly lower, we still set the lower cutoff to 4 due to more eliminated terms and reduced computational time. Additionally, there is no clear pattern indicating the amount of terms that gives the highest prediction accuracy. Including more terms does not necessarily improve accuracy due to the negative effects of noise words and stop words, and including fewer terms might lead to loss of information, consequently the need for an efficient term selection method is emphasized, namely Zipf's law.

6.1.3. Experiment C. Performance comparison against other classification techniques

Concerning performance of the proposed model, experiments were conducted to evaluate prediction accuracy against other traditional classification techniques using Waikato Environment for Knowledge Analysis (WEKA) 3.4 [35]. WEKA contains a large number of classifiers divided into several groups, namely Bayesian, based on functions, lazy, meta-techniques, tree-based, etc. Among each group, classifiers BayesNet (BN), NaiveBayes (NB), NaiveBayesSimple (NBS), MultilayerPerceptron (MLP), KStar, Bagging and decision tree J48 were chosen and tested for different membership degree thresholds and two sets of upper and lower cutoffs. One-tailed paired *t*-tests on the proposed ICD-AS method against these traditional methods are conducted to determine significancy of the comparisons, respectively.

Table 7 shows prediction accuracy of the proposed model and seven traditional classifiers for upper cutoff set to 1, lower cutoff set to 1, and membership degree ranging from 0.00 to 0.50; and the *p*-values of the one-tailed *t*-tests at 95% confidence level are listed in the last row of the table. This combination of upper and lower cutoffs means that all words are considered and no stop-words nor noise-words are eliminated, not taking into account the importance of attribute selection. Highlighted figures indicate the best values obtained on each membership degree threshold. Our model ICD-AS achieves the highest prediction accuracy with almost all membership degree thresholds, is slightly less accurate when membership degree threshold is set to 0.00 and 0.10, and has the best overall performance with an average accuracy of 76.23%. The one-tailed *t*-tests also show that the performance improvement of the proposed ICD-AS model is statistically significant. In addition, only ICD-AS, MPL and KStar present logical accuracy trends because their highest prediction accuracies are for membership degree thresholds larger than 0.00.

Table 8 shows prediction accuracy of ICD-AS and seven traditional classifiers for upper cutoff set to 3, lower cutoff set to 4, and membership degree threshold ranging from 0.00 to 0.50; and the *p*-values of the one-tailed *t*-tests at 95% confidence level are listed in the last row of the table. This combination of upper and lower cutoffs is the most suitable as it yields the highest overall average accuracy (refer to Section 5.3.2). Our model ICD-AS achieves the highest prediction accuracy with almost all membership degree thresholds, is slightly less accurate when membership degree threshold is set to 0.00, and has the best overall performance with an average accuracy of 76.77%. Again, the *p*-values also show that the proposed ICD-AS model improves significantly on prediction accuracy.

Evaluation of the system reveals that it performs ICD code prediction relatively well, indicating that disease classification specialists might benefit from it by considering the code suggested,

Table 6
Average prediction accuracy for different combinations of upper and lower cutoffs.

Upper cutoff	Lower cutoff								Average accuracy
	1	2	3	4	5	6	7	8	
1	76.23%	76.06%	74.95%	75.37%	74.90%	75.58%	73.12%	72.43%	74.83%
2	75.37%	76.98%	74.69%	75.28%	74.82%	75.92%	73.54%	72.43%	74.88%
3	75.67%	75.93%	75.59%	76.77%	75.49%	76.00%	74.00%	74.04%	75.44%
4	73.47%	73.76%	73.10%	75.34%	73.70%	73.53%	70.91%	71.26%	73.14%
5	75.64%	74.02%	73.09%	74.82%	73.53%	73.36%	71.25%	71.78%	73.44%
6	75.90%	74.02%	72.97%	74.94%	73.88%	73.79%	71.85%	72.56%	73.74%
Average accuracy	75.38%	75.13%	74.07%	75.42%	74.39%	74.70%	72.45%	72.42%	

Table 7
Prediction accuracies with upper cutoff set to 1 and lower cutoff set to 1.

Membership degree threshold	ICD-AS	BN	NB	NBS	MLP	KStar	Bagging	J48
0.00	80.34%	80.00%	80.00%	80.00%	81.11%	76.67%	80.00%	80.28%
0.10	81.20%	80.00%	80.28%	80.28%	82.50%	79.17%	79.17%	80.00%
0.15	82.05%	77.50%	77.78%	77.78%	77.22%	76.39%	76.11%	77.50%
0.20	82.05%	76.39%	76.11%	76.11%	59.72%	71.39%	75.28%	75.56%
0.25	80.34%	74.17%	73.33%	73.33%	54.16%	66.67%	72.22%	70.00%
0.30	78.63%	70.00%	68.33%	68.33%	49.72%	62.78%	68.33%	69.17%
0.35	74.36%	68.33%	66.39%	66.39%	48.61%	60.56%	66.67%	66.94%
0.40	70.59%	64.72%	62.78%	62.78%	47.78%	56.94%	61.11%	56.39%
0.45	69.75%	61.39%	60.28%	60.28%	46.94%	57.22%	58.89%	58.61%
0.50	63.03%	57.22%	56.94%	56.94%	49.44%	53.06%	58.33%	56.11%
Average	76.23%	70.97%	70.22%	70.22%	59.72%	66.09%	69.61%	69.06%
p-Value		7.76E−5	1.44E−4	1.44E−4	7.34E−4	4.26E−5	9.50E−5	2.97E−4

Table 8
Prediction accuracies with upper cutoff set to 3 and lower cutoff set to 4.

Membership degree threshold	ICD-AS	BN	NB	NBS	MLP	KStar	Bagging	J48
0.00	80.51%	81.56%	79.89%	79.89%	80.17%	79.61%	79.61%	80.17%
0.10	82.20%	80.73%	79.61%	79.61%	81.56%	79.89%	79.05%	79.89%
0.15	81.51%	78.49%	77.37%	77.37%	78.77%	75.98%	77.37%	79.05%
0.20	78.99%	75.42%	74.86%	74.86%	72.91%	69.83%	74.58%	75.98%
0.25	75.21%	70.95%	71.51%	71.51%	67.60%	65.08%	70.39%	70.39%
0.30	77.12%	68.44%	67.88%	67.88%	66.76%	61.45%	66.76%	65.36%
0.35	77.97%	65.36%	65.64%	65.64%	60.89%	58.66%	63.13%	62.01%
0.40	75.00%	60.89%	60.06%	60.06%	60.61%	57.26%	58.94%	58.94%
0.45	74.79%	57.82%	56.98%	56.98%	56.98%	55.59%	57.82%	56.70%
0.50	64.41%	52.23%	52.51%	52.51%	54.19%	51.11%	53.63%	56.15%
Average	76.77%	69.19%	68.63%	68.63%	68.04%	65.45%	68.13%	68.46%
p-Value		1.78E−3	8.94E−4	8.94E−4	9.82E−4	2.65E−4	6.16E−4	1.70E−3

which is strengthened by a certainty factor increasing its credibility. Moreover, implication rules calculated also help to uncover hidden knowledge that might have been omitted by users.

6.2. Discharge summary set 2

In this experiment, we conduct a performance comparison with a linear-kernel SVM [36] using a larger set consisting of 2579 discharge summaries. The academic tool Weka LibSVM (WLSVM) is used for the comparison, which integrates the well-recognized LibSVM into Weka Environment [37]. The distribution of the set is imbalanced among five ICD codes, namely 430, 431, 432, 433, and 434, of each contains 99, 446, 22, 56, and 1956 discharge summaries, respectively. Following the same procedure of discharge summary set 1, two major upper and lower cutoff sets are empirically determined as (40, 4) and (10, 4), of each 336 and 360 terms were identified by the MeSH database as root terms, respectively. This experiment performs 10-fold cross validation with appropriate parameter setting, and the comparison results are made in terms of precision, recall, *F*-measure and AUC, the area under

Table 9
Performance compassion of ICD-AS and linear-kernel SVM with upper cutoff = 40 and lower cutoff = 4.

	ICD-AS				Linear-kernel SVM			
	Precision	Recall	*F*-measure	AUC	Precision	Recall	*F*-measure	AUC
430	63.00%	42.60%	50.86%	79.10%	76.30%	44.60%	56.30%	72.00%
431	66.50%	51.20%	57.83%	78.70%	67.40%	47.30%	55.60%	71.50%
432	68.40%	77.30%	72.57%	98.10%	63.30%	86.40%	73.10%	92.80%
433	42.60%	22.20%	29.21%	72.90%	59.40%	38.00%	46.30%	68.60%
434	87.10%	94.00%	90.45%	81.60%	86.10%	94.20%	90.00%	73.80%
Weighted average	81.40%	83.00%	82.16%	80.90%	81.60%	82.80%	81.50%	73.40%
p-Value					3.05E−1	2.54E−1	3.08E−1	6.54E−4

Table 10
Performance compassion of ICD-AS and linear-kernel SVM with upper cutoff = 10 and lower cutoff = 4.

	ICD-AS				Linear-kernel SVM			
	Precision	Recall	*F*-measure	AUC	Precision	Recall	*F*-measure	AUC
430	87.60%	76.30%	81.54%	97.30%	80.70%	88.90%	84.60%	94.00%
431	85.70%	71.80%	78.13%	90.50%	87.70%	74.40%	80.50%	86.10%
432	94.50%	72.70%	82.21%	96.00%	64.30%	81.80%	72.00%	90.70%
433	64.30%	21.40%	32.14%	74.00%	61.50%	42.90%	50.50%	71.10%
434	91.40%	97.60%	94.37%	91.20%	93.10%	96.40%	94.70%	87.30%
Weighted average	89.60%	90.20%	89.89%	90.95%	90.70%	90.90%	90.60%	87.10%
p-Value					2.92E−1	8.76E−2	5.76E−1	7.15E−4

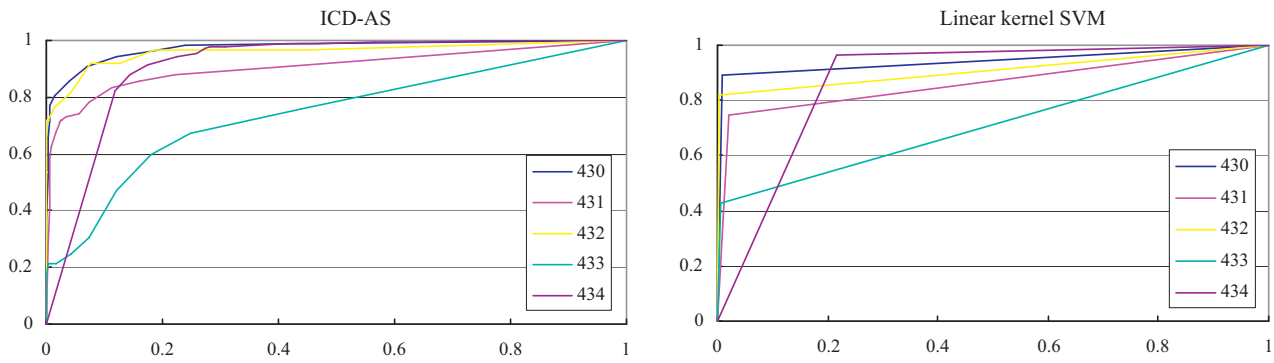

Fig. 10. The ROC curve comparison of ICD-AS and linear-kernel SVM.

the ROC (receiver operating characteristic) curve, a widely used performance measure of a classification model. Tables 9 and 10 summarize the comparison on the basis of two different cutoff values. Two-tailed paired *t*-tests are conducted to determine the significance of the difference between the two models on precision, recall, *F*-measure and AUC, respectively. The *p*-values for 95% confidence level are shown in the last rows of both tables, which indicate that in both experiments, the proposed ICD-AS achieves higher AUC values significantly, and that it performs as good as the linear-kernel SVM on precision, recall, and *F*-measure at 95% confidence level.

Fig. 10 illustrates the ROC curve comparison of ICD-AS with upper cutoff = 10 and lower cutoff = 4, and the linear-kernel SVM for five ICD codes. It demonstrates that both classification models achieve the prediction performance consistently in the order of codes 430, 432, 431, 434 and 433.

7. Conclusion and future work

Nowadays, with the blooming of the knowledge era, knowledge generation and sharing has become essential in every organization, but has also lead to severe data overflow. Thus, it is necessary to find efficient ways to manage the abundant knowledge produced and turn it useful. In this paper we provided a semi-automatic method based on FFCA that accurately suggests ICD codes for discharge summaries and performs knowledge extraction through implication rules. Under our research framework, disease classification specialists can achieve a better understanding of the relationships between medical terms and be supported in the coding of discharge summaries, reducing the possibility of error that may decrease monetary income for hospitals.

Discharge summaries of patients suffering from cerebrovascular disease were processed using information retrieval techniques to identify medical terms, and analysis following Zipf's law was applied to adequately perform attribute selection by approximating upper and lower cutoffs that helped distinguish significant terms. A membership degree threshold was established to eliminate relations between terms and discharge summaries with low significance, and similarity values between concepts were calculated to allow prediction of ICD codes based on a voting scheme. Furthermore, a strength value was computed for each possible ICD code to determine the most suitable to be suggested for a discharge summary of interest, and a certainty factor representing a degree of belief was assessed to support each ICD code prediction. Performance evaluation of the proposed system was carried out by testing the prediction of ICD codes for a set of discharge summaries and experimental results showed that the method proposed successfully accomplishes clustering of discharge summaries based on content similarity and correctly suggests ICD codes. The disease classification method in the proposed ICD-AS can be realized by traditional classifiers such as decision trees, naïve Bayes classifier, SVM and so on. Our method differs from them in the nature of conceptual clustering powered by FFCA since FCA can be used for discovering inherent relationships between objects described through a set of attributes on the one hand, and the attributes themselves on the other [38]. In the sense, it not only aims at determining clusters, but provides at the same time intensional descriptions of these extensions [39]. Therefore, according to the experimental comparison, ICD-AS achieves better performance than the traditional classifiers. More importantly, the system properly produces formal concepts and implication rules, which improve comprehension of the field of cerebrovascular disease and give insights to the relationships between relevant medical terms. On the other hand, the major shortcoming of ICD-AS lies in the efficiency issue of analyzing the formal concept lattice when the size of the formal context, i.e. the number of discharge summaries, increases significantly since FCA is not suited for direct manipulation of very large data sources, as addressed by [2].

The large demand for healthcare in Taiwan deems optimal medical resource management a must in order to provide outstanding cheap healthcare for all. Our system contributes valuable guidance to disease classification specialists in the process of coding discharge summaries, which consequently brings benefits in many aspects of society. At the patient level, better and more adequate treatment will be received if hospitals have sufficient resources and clear procedures established, reducing patients' length of stay in hospitals and increase quality of treatment. At the hospital level, junior disease classification specialists will be better supported during the coding of discharge summaries and senior staff will be able to share their knowledge more easily, improving the overall quality of the coding process. In addition, hospitals will use resources more efficiently and reduce waste, decreasing their economic loss and improving the quality of healthcare provided. At the healthcare system level, the BNHI will have better control over its resources, allocate limited resources more evenly and reduce overall costs. It will also be able to measure national clinical activity more closely and improve assessment and planning of healthcare activities.

Due to the large amount of combinations of available ICD codes, our work focused on a limited range belonging to the area of cerebrovascular disease; hence a more extensive analysis should be made using a wider range of codes that would provide a more complete coverage under an acceptable efficiency consideration aforementioned. The input documents are restricted to clinical related documents, in this case English hospital discharge summaries, due to the domain of the dictionary used for term filtering. The content of the discharge summaries must be complete and accurate so that the correct knowledge can be properly elicited and their length is also a significant factor as scalability is one of the major hindrances to formal concept analysis. Future work

should also be aimed at employing ontologies to represent lattice diagrams such that knowledge exploration can be integrated with other disciplines. Ontological engineers and specialists from other fields would be able to view the knowledge structure from different perspectives and it could be extended and used in a broader variety of applications. Finally, our framework should be better integrated in order to make the system entirely automatic, enhancing its usability and functionality.

Conflicts of interest

None.

Acknowledgements

This study was supported in part by National Science Council NSC98-2410-H-006-007, Taiwan, ROC. The authors thank Prof. Chih-Ping Wei for his useful suggestion on preparing the draft and also appreciate Dr. Wei Rong Zhou, M.D. for providing his expertise in cerebrovascular disease to help the validation of implication rules extracted.

References

[1] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival I, editor. Ordered sets. Dordrecht: Reidel; 1982. p. 445–70.
[2] Priss U. Formal concept analysis in information science. In: Cronin B, editor. Annual review of information science and technology. Medford: Information Today, Inc.; 2006. p. 521–43.
[3] Formica A. Ontology-based concept similarity in formal concept analysis. Information Sciences 2006;176:2624–41.
[4] Wolff KE. A first course in formal concept analysis - how to understand line diagrams. In: Faulbaum F, editor. 7th conference on the scientific use of statistical software. Stuttgart: Gustav Fischer Verlag; 1993. p. 429–38.
[5] Chu S, Cesnik B. Knowledge representation and retrieval using conceptual graphs and free text document self-organisation techniques. International Journal of Medical Informatics 2001;62:121–33.
[6] Diaz-Agudo B, Gonzalez-Calero PA. Formal concept analysis as a support technique for CBR. Knowledge-Based Systems 2001;14:163–71.
[7] Jiang G, Ogasawara K, Endoh A, Sakurai T. Context-based ontology building support in clinical domains using formal concept analysis. International Journal of Medical Informatics 2003;71:71–81.
[8] Boucher-Ryan PD, Bridge D. Collaborative recommending using formal concept analysis. Knowledge-Based Systems 2006;19:309–15.
[9] Carpineto C, Romano G. Exploiting the potential of concept lattices for information retrieval with CREDO. Journal of Universal Computer Science 2004;10:985–1013.
[10] Tho QT, Hui SC, Fong ACM. A citation-based document retrieval system for finding research expertise. Information Processing and Management 2007;43:248–64.
[11] Rajapakse RK, Denham M. Text retrieval with more realistic concept matching and reinforcement learning. Information Processing and Management 2006;42:1260–75.
[12] Weng S-S, Tsai H-J, Liu S-C, Hsu C-H. Ontology construction for information classification. Expert Systems with Applications 2006;31:1–12.
[13] Tho QT, Hui SC, Cao TH. A fuzzy FCA approach for citation-based document retrieval. In: Ge S, Tan K, editors. Proceedings of the 2004 IEEE conference on cybernetics and intelligent systems. Singapore: IEEE Press; 2004. p. 578–83.
[14] Belohlavek R, Dvorak J, Outrata J. Fast factorization by similarity in formal concept analysis of data with fuzzy attributes. Journal of Computer and System Sciences 2007;73:1012–22.
[15] Jiang G, Pathaka J, Christopher GC. Formalizing ICD coding rules using formal concept analysis. Journal of Biomedical Informatics 2009;42:504–17.
[16] Jay N, Napoli A, Kohler F. Cancer patient flows discovery in DRG databases. Studies in Health Technology and Informatics 2006;124:725–30.
[17] Tho QT, Hui SC, Fong ACM, Cao TH. Automatic fuzzy ontology generation for semantic web. IEEE Transactions on Data and Knowledge Engineering 2006;18:842–56.
[18] Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. Harlow: Pearson Education Limited; 1999.
[19] Zipf GK, Selective. Studies and the principle of relative frequency in language. Cambridge: MIT Press; 1932.
[20] Luhn HP. The automatic creation of literature abstracts. IBM Journal of Research and Development 1958;2:157–65.
[21] Adamic LA, Huberman BA. Zipf's law and the Internet. Glottometrics 2002;3:143–50.
[22] Krashakov SA, Teslyuk AB, Shchur LN. On the universality of rank distributions of website popularity. Computer Networks 2006;50:1769–80.
[23] Kalda J, Sakki M, Vainu M, Laan M. Zipf's law in human heartbeat dynamics, http://arxiv.org/PS_cache/physics/pdf/0110/0110075v1.pdf; 2009 [accessed 27.03.09].
[24] Salton G, McGill MJ. Introduction to modern information retrieval. New York: McGraw Hill Publishing Company; 1983.
[25] Leroy G, Chen H. Introduction to the special issue on decision support in medicine. Decision Support Systems 2007;43:1203–6.
[26] Greengrass E. Information retrieval: a survey, DOD technical report TR-R52-008-001, Baltimore, UMBC; 2001, 224 pp.
[27] Houston A, Chen H, Schatz BR, Hubbard SM, Sewell RR, Ng TD. Exploring the use of concept spaces to improve medical information retrieval. Decision Support Systems 2000;30:171–86.
[28] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing and Management 1988;24:513–23.
[29] Shortliffe EH, Buchanan BG. A model of inexact reasoning in medicine. Mathematical Biosciences 1975;23:351–79.
[30] Ishibuchi H, Nozaki K, Yamamoto N, Tanaka H. Selecting fuzzy if-then rules for classification problems using genetic algorithms. IEEE Transactions on Fuzzy Systems 1995;3:260–70.
[31] Yevtushenko SA. System of data analysis "Concept Explorer". In: Proceedings of the 7th National Conference on Artificial Intelligence KII-2000. Russia, Moscow: ACM; 2000. p. 127–34.
[32] Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intelligent Data Analysis 2002;6:429–49.
[33] Hulse JV, Khoshgoftaar TM, Napolitano A. Experimental perspectives on learning from imbalanced data. In: Ghahramani Z, editor. Proceedings of the 24th international conference on machine learning. Corvalis: International Machine Learning Society; 2007. p. 935–42.
[34] Green R, Bean CA, Myaeng SH, editors. The semantics of relationships: an interdisciplinary perspective. Boston: Kluwer Academic Publishers; 2002.
[35] Garner SR. WEKA: the waikato environment for knowledge analysis. In: Reeves S, Cranefield S, editors. Proceedings of the second New Zealand computer science research students conference. Hamilton: University of Waikato; 1995. p. 57–64.
[36] Joachims T. Learning to classify text using support vector machines methods, theory, and algorithms. Dordrecht: The Netherlands Kluwer Academic Publishers; 2002.
[37] EL-Manzalawy Y, Honavar V. WLSVM: integrating LibSVM into weka environment, http://www.cs.iastate.edu/~yasser/wlsvm; 2010 [accessed 05.06.10].
[38] Ganter B, Wille R. Formal concept analysis: mathematical foundations. Berlin: Springer Verlag; 1999.
[39] Cole R, Lund P. Document retrieval for e-mail search and discovery using formal concept analysis. Applied Artificial Intelligence 2003;17:257–80.