Bridging the Vocabulary Gap between Health Seekers and Healthcare Knowledge

Liqiang Nie, *Member, IEEE*, Yi-Liang Zhao, Mohammad Akbari, Jialie Shen, *Member, IEEE*, and Tat-Seng Chua, *Member, IEEE*

Abstract—The vocabulary gap between health seekers and providers has hindered the cross-system operability and the inter-user reusability. To bridge this gap, this paper presents a novel scheme to code the medical records by jointly utilizing local mining and global learning approaches, which are tightly linked and mutually reinforced. Local mining attempts to code the individual medical record by independently extracting the medical concepts from the medical record itself and then mapping them to authenticated terminologies. A corpus-aware terminology vocabulary is naturally constructed as a byproduct, which is used as the terminology space for global learning. Local mining approach, however, may suffer from information loss and lower precision, which are caused by the absence of key medical concepts and the presence of irrelevant medical concepts. Global learning, on the other hand, works towards enhancing the local medical coding via collaboratively discovering missing key terminologies and keeping off the irrelevant terminologies by analyzing the social neighbors. Comprehensive experiments well validate the proposed scheme and each of its component. Practically, this unsupervised scheme holds potential to large-scale data.

Index Terms—Healthcare, medical terminology assignment, global learning, local mining, question answering

1 Introduction

Information technologies are transforming the ways healthcare services are delivered, from patients' passively embracing their doctors' orders to patients' actively seeking online information that concerns their health. This trend is further confirmed by a national survey conducted by the Pew Research Center[1] in Jan 2013, where they reported that one in three American adults have gone online to figure out their medical conditions in the past 12 months from the report time.

To better cater to health seekers, a growing number of community-based healthcare services have turned up, including HealthTap,[2] HaoDF[3] and WebMD.[4] They are disseminating personalized health knowledge and connecting patients with doctors worldwide via question answering [1], [2]. These forums are very attractive to both professionals and health seekers. For professionals, they are able to increase their reputations among their colleagues and patients, strengthen their practical knowledge from interactions with other renowned doctors, as well as possibly attract more new patients. For patients, these systems provide nearly instant and trusted answers especially for complex and sophisticated problems. Over times, a tremendous number of medical records have been accumulated in their repositories, and in most circumstances, users may directly locate good answers by searching from these record archives, rather than waiting for the experts' responses or browsing through a list of potentially relevant documents from the Web.

In many cases, the community generated content, however, may not be directly usable due to the vocabulary gap. Users with diverse backgrounds do not necessarily share the same vocabulary. Take HealthTap as an example, which is a question answering site for participants to ask and answer health-related questions. The questions are written by patients in narrative language. The same question may be described in substantially different ways by two individual health seekers. On the other side, the answers provided by the well-trained experts may contain acronyms with multiple possible meanings, and non-standardized terms. Recently, some sites have encouraged experts to annotate the medical records with medical concepts. However, the tags used often vary wildly and medical concepts may not be medical terminologies [3]. For example, "heart attack" and "myocardial disorder" are employed by different experts to refer to the same medical diagnosis. It was shown that the inconsistency of community generated health data greatly hindered data exchange, management and integrity [4]. Even worse, it was reported that users had encountered big challenges in reusing the archived content due to the incompatibility between their search terms and those accumulated medical records [5]. Therefore, automatically coding the medical records with standardized terminologies is highly desired. It leads to a consistent interoperable way

[1] http://pewinternet.org/Reports/2013/Health-online.aspx.
[2] https://www.healthtap.com/.
[3] http://www.haodf.com/.
[4] http://www.webmd.com/.

- *L. Nie, Y.-L. Zhao, M. Akbari, and T.-S. Chua are with the School of Computing, National University of Singapore, Singapore 117417. E-mail: {nieliqiang, zhaoyiliang}@gmail.com, akbari@nus.edu.sg, chuats@comp.nus.edu.sg.*
- *J. Shen is with the School of Information Systems, Singapore Management University, Singapore. E-mail: jlshen@smu.edu.sg.*

Manuscript received 4 Sept. 2013; revised 30 May 2014; accepted 30 May 2014. Date of publication 24 June 2014; date of current version 23 Dec. 2014.
Recommended for acceptance by A. Zhang.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2014.2330813

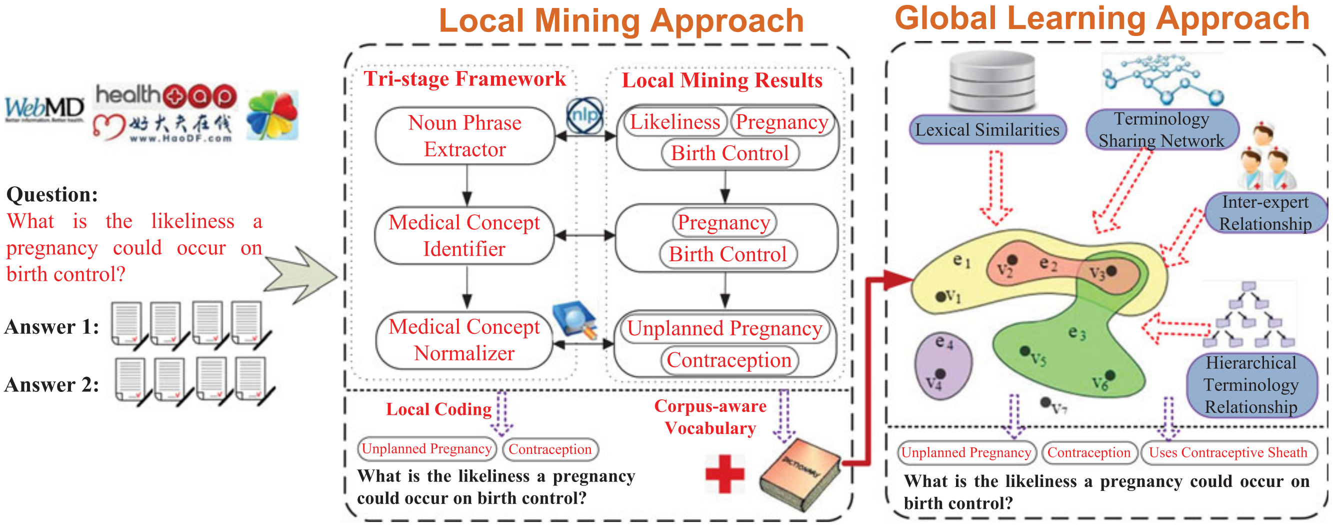

Fig. 1. The schematic illustration of the proposed automatic medical terminology assignment scheme. The answer part is not displayed due to the space limitation.

of indexing, storing and aggregating across specialties and sites. In addition, it facilitates the medical record retrieval via bridging the vocabulary gap between queries and archives.

It is worth mentioning that there already exist several efforts dedicated to research on automatically mapping medical records to terminologies [6], [7], [8], [9], [10], [11]. Most of these efforts, however, focused on hospital generated health data or health provider released sources by utilizing either isolated or loosely coupled rule-based and machine learning approaches. Compared to these kinds of data, the emerging community generated health data is more colloquial, in terms of inconsistency, complexity and ambiguity, which pose challenges for data access and analytics. Further, most of the previous work simply utilizes the external medical dictionary to code the medical records rather than considering the corpus-aware terminologies. Their reliance on the independent external knowledge may bring in inappropriate terminologies. Constructing a corpus-aware terminology vocabulary to prune the irrelevant terminologies of specific dataset and narrow down the candidates is the tough issue we are facing. In addition, the varieties of heterogeneous cues were often not adequately exploited simultaneously. Therefore, a robust integrated framework to draw the strengths from various resources and models is still expected.

We propose a novel scheme that is able to code the medical records with corpus-aware terminologies. As illustrated in Fig. 1, the proposed scheme consists of two mutually reinforced components, namely, local mining and global learning. Local mining aims to locally code the medical records by extracting the medical concepts from individual record and then mapping them to terminologies based on the external authenticated vocabularies. We establish a tri-stage framework to accomplish this task, which includes noun phrase extraction, medical concept detection and medical concept normalization. As a byproduct, a corpus-aware terminology vocabulary is naturally constructed, which can be used as terminology space for further learning in the second component. However, local mining approach may suffer from the problem of information loss and low precision due to the possible lack of some key medical concepts in the medical records and the presence of some irrelevant medical concepts. We thus propose global learning to complement the local medical coding in a graph-based approach. It collaboratively learns missing key concepts and propagates precise terminologies among underlying connected records over a large collection. Besides the semantic similarity among medical records and terminology-sharing network, the inter-terminology and inter-expert relationships are seamlessly integrated in the proposed model. The inter-terminology relationships are mined by exploiting the external well-structured ontology, which are able to alleviate the granularity mismatch problems and reduce the irrelevant sibling terminologies. The inter-expert relationships are inferred from the experts' historical data. It may be capable of excluding a wealth of domain-specific context information. Specifically, the medical professionals who are frequently respond to the same kinds of questions probably share highly overlapping expertise, and thus the questions they answered can be regarded as semantically similar to a certain extent. Extensive evaluations on the real-world dataset demonstrate that our proposed scheme can achieve significant gains in medical terminology assignment. Meanwhile, the whole process of our proposed approach is unsupervised and it holds potential to handle large-scale data.

The main contributions of this work are threefold:

- To the best of our knowledge, this is the first work on automatically coding the community generated health data, which is more complex, inconsistent and ambiguous compared to the hospital generated health data.
- It proposes the concept entropy impurity (CEI) approach to comparatively detect and normalize the medical concepts locally, which naturally construct a corpus-aware terminology vocabulary with the help of external knowledge.

- It builds a novel global learning model to collaboratively enhance the local coding results. This model seamlessly integrates various heterogeneous information cues.

The remainders are structured as follows. Section 2 briefly reviews the related work. The local mining and global learning approaches are respectively introduced in Sections 3 and 4. Section 5 details the experimental results and analysis, followed by our concluding remarks in Section 7.

2 Related Work

Most of the current health providers organize and code the medical records manually [3]. This workflow is extremely expensive because only well-trained experts are properly competent for the task. Therefore, there is a growing interest to develop automated approaches for medical terminology assignment. The existing techniques can be categorized into two categories: rule-based and machine learning approaches.

Rule-based approaches play a principle role in medical terminology assignments [6], [7], [8]. They generally discover and construct effective rules by making strong uses of the morphological, syntactic, semantic and pragmatic aspects of natural language. It has been found that these methods have significant positive effects on the real systems [12]. Back in 1995, Hersh and David [13] designed and developed a system, named SAPPHIRE, which automatically assigned UMLS[5] terminologies to medical documents using a simple lexical approach. Around one decade later, a system named IndexFinder [14], proposed a new algorithm for generating all valid UMLS terminologies by permuting the set of words in the input text and then filtering out the irrelevant concepts via syntactic and semantic filtering. Most recently, several efforts [12], [15], [16], [17] have attempted to automatically convert free medical texts into medical terminologies/ontologies by combining several natural language processing methods, such as stemming, morphological analysis, lexicon augmentation, term composition and negation detection. However, these methods are purely applicable to well-constructed discourses. A proposal in [4], instead of just converting the corpus data to terminologies, suggested users with appropriate medical terminologies for their personal queries. It integrated UMLS, WordNet as well as Noun Phraser to capture the semantic meaning of the queries. However, an implicit assumption of this work is that the sources to be searched must be well presented using a standardized medical vocabulary. Obviously, this is not applicable to the community generated medical sources. In summary, even though rule-based methods are fast and suitable for real-time applications, the rule construction is challenging and the performance varies from different corpus.

Machine learning approaches build inference models from medical data with known annotations and then apply the trained models to unseen data for terminology prediction [6], [18]. The research can be traced back to the 1990 s, where Larkey and Croft [10] have trained three statistical classifiers and combined their results to obtain a better classification in 1995. In the same year, support vector machine (SVM) and Bayesian ridge regression were first evaluated on large-scale dataset and obtained promising performance [9]. Following that, a hierarchical model was studied in [19], which exploited the structure of ICD-9 code set and demonstrated that their approach outperformed the algorithms based on the classic vector space model. About ten years later, Suominen et al. [11] introduced a cascade of two classifiers to assign diagnostic terminologies to radiology reports. In their model, when the first classifier made a known error, the output of the second classifier was used instead to give the final prediction. Yan et al. [20] proposed a multi-label large-margin formulation that explicitly incorporated the inter-terminology structure and prior domain knowledge simultaneously. This approach is feasible for small terminology set but is questionable in real-life settings where thousands of terminologies need to be considered.

Similar to our scheme, Pakhomov et al. [21] attempted to improve the coding performance by combing the advantages of rule-based and machine learning approaches. It described Autocoder, an automatic encoding system implemented at Mayo clinic. Autocoder combines example-based rules and a machine learning module using Naïve Bayes. However, this integration is loosely coupled and the learning model can not incorporate heterogeneous cues, which is not a good choice for the community-based health services.

Beyond medical domain, several prior efforts of corpus alignment and gap bridging have been dedicated to other verticals. Chen et al. [22] derived an integrated model that jointly aligns bilingual named entities between Chinese and English news. The work in [23] bridged the management research-practice gap by describing their experiences with the network for business sustainability. A game platform was designed in [24] and was demonstrated how to enhance the inter-generation cultural communication in a family. These diverse efforts are all heuristic. Their rules and patterns are domain specific and cannot be generalized to other areas. Another example, the music semantic gap between textual query and audio content was remedied by annotation with concepts [25]. This approach can hardly be applied to medical terminology assignment directly due to the differences in modalities and content structures. Besides, it targets at labeling music entities with common noun and adjective phrases, while our approach focuses on terminologies only.

3 Local Mining

Medical concepts are defined as medical domain-specific noun phrases, and medical terminologies are referred to as authenticated phrases by well-known organizations that are used to accurately describe the human body and associated components, conditions and processes in a science-based manner. This section details the local mining approach. To accomplish this task, we establish a tri-stage framework. Specifically, given a medical record, we first extract the embedded noun phrases. We then identify the medical concepts from these noun phrases by measuring their specificity. Finally, we normalize the detected medical concepts to terminologies.

[5] http://www.nlm.nih.gov/research/umls/.

3.1 Noun Phrase Extraction

To extract all the noun phrases, we initially assign part-of-speech tags to each word in the given medical record by Stanford POS tagger.[6] We then pull out sequences that match a fixed pattern as noun phrases. This pattern is formulated as follows:

$$\begin{aligned}&(Adjective|Noun)^*(Noun \quad Preposition)\\&?(Adjective|Noun)^*Noun.\end{aligned} \tag{1}$$

The above regular expression can be intuitively interpreted as follows. The noun phrases should contain zero or more adjectives or nouns, followed by an optional group of a noun and a preposition, followed again by zero or more adjectives or nouns, followed by a single noun. A sequence of tags matching this pattern ensures that the corresponding words make up a noun phrase. For example, the following complex sequence can be extracted as a noun phrase: "ineffective treatment of terminal lung cancer". In addition to simply pulling out the phrases, we also do some simple post processing to link the variants together, such as singularizing plural variants.

3.2 Medical Concept Detection

This stage aims to differentiate the medical concepts from other general noun phrases. Inspired by the efforts in [26], we assume that concepts that are relevant to medical domain occur frequently in medical domain and rarely in non-medical ones. Based on this assumption, we employ the concept entropy impurity [26] to comparatively measure the domain-relevance of a concept. For a concept c, its CEI is computed as follows:

$$CEI(c) = -\sum_{i=1}^{2} P(D_i|c)logP(D_i|c), \tag{2}$$

where D_1 and D_2 respectively represents our medical corpus and a general-domain corpus; and $P(D_i|c)$ denotes the probability that a concept c is related to a specified domain D_i; $P(D_i|c)$ can be computed as

$$P(D_i|c) = \frac{count(c, D_i)}{count(c)}. \tag{3}$$

To remove the effect of different corpus's length, we define the normalized $P_n(D_i|c)$ as follows:

$$P_n(D_i|c) = \frac{P(D_i|c)/L_i}{\sum_{j=1}^{2} P(D_j|c)/L_j}, \tag{4}$$

where L_i is the sum of document lengths in D_i. Obviously, $CEI(c)$ reaches the maximum value of 0.693, when concept c equally distributes within these two corpus. This implies that the larger CEI of a concept is, the more domain-irrelevant is it. To make it easily computer-processable, we define specificity of a concept to the medical domain as follows:

$$specificity(c) = \begin{cases} 1 - \alpha CEI(c), & \text{if } P_n(D_1|c) > P_n(D_2|c),\\ \alpha CEI(c), & \text{otherwise,}\end{cases} \tag{5}$$

where $\alpha = \frac{0.5}{0.693}$. Meanwhile, a threshold is set to detect the medical concepts.

3.3 Medical Concept Normalization

Although medical concepts are defined as medical domain-specific noun phrases, we cannot ensure that they are standardized terminologies. Take "birth control" as an example, it is recognized as a medical concept by our approach, but it is not an authenticated terminology. Instead, we should map it into "contraception". Therefore, it is essential to normalize the detected medical concepts according to the external suitable standardized dictionary and this normalization is the key to bridging the vocabulary gap.

Currently, there exist numerous authenticated vocabularies, including ICD,[7] UMLS, and SNOMED CT.[8] These medical and clinical terminologies were created in different times by different associations for different purposes. Take ICD as an example: it is typically used for external reporting requirements or other uses where data aggregation is advantageous. In this work, we use SNOMED CT because it provides the core general terminologies for the electronic health record and formal logic-based hierarchical structure.

The terminologies and their descriptions in SNOMED CT are first indexed.[9] We then search each medical concept against the indexed SNOMED CT. For the medical concepts with multiple matched results, e.g., two results returned for "female", we keep all the returned terminology candidates (i.e., fully specified concept) for further selection. Enlightened by Google distance [27] that is concepts with the same or similar meanings in a natural language sense tend to be "close" in units of Google distance, while concepts with dissimilar meanings tend to be farther apart, we estimate the semantic similarity between the medical concept and the returned terminology candidates via exploring their co-occurrence on Google:

$$d(t_i, c) = \frac{\max(\log r(t_i), \log r(c)) - \log r(t_i, c)}{\log G - \min(\log r(t_i), \log r(c))}, \tag{6}$$

where G is the total number of documents retrieved from Google; t_i and c respectively represents the terminology candidate and the medical concept; $r(x)$ is the number of hits for search concepts x; and $r(t_i, c)$ is the number of web documents in which both t_i and c co-occur. Then their semantic relevance is defined as:

$$S(t_i, c) = \exp(-d(t_i, c)). \tag{7}$$

We then select the most relevant terminology candidate as the normalized result.

[6] http://nlp.stanford.edu/software/tagger.shtml.

[7] http://www.who.int/classifications/icd/en/.

[8] http://www.ihtsdo.org/snomed-ct/.

[9] http://viw2.vetmed.vt.edu/sct/menu.cfm.

3.4 Discussions

Each medical record is coded with multiple terminologies with local mining, which are generated via mapping their embedded medical concepts to terminologies. However, these mined terminologies may suffer from various problems.

The first problem is incompleteness. This is because some key medical concepts not explicitly present in the medical records are excluded. The medical record illustrated in Fig. 1 shows such an example, where the accurate terminology: "use contraceptive sheath" is absent from the medical record.

The second one is the lower precision. This is due to some irrelevant medical concepts explicitly embed in the medical records, and are mistakenly detected and normalized by the local approach. Take the second medical record in Table 6 as an example, where "finding of life event" normalized from irrelevant medical concept "life" is assigned as code, even though it is less informative to capture the main intent.

Another issue, which deserves further discussion here, is the terminology space. Most previous efforts, including our local approach, attempted to map the medical records directly to the entries in external dictionaries without any pruning. There is a problem to do so since the external dictionaries usually cover relatively comprehensive terminologies and are far beyond the vocabulary scope of the given corpus. It may result in the deterioration in coding performance in terms of efficiency and effectiveness. The problem is caused by the over-widened scope of vocabularies, which may bring in unpredictable noises and make the precise terminology selection challenging.

4 Graph-Based Global Learning

Let $\mathcal{Q} = \{q_1, q_2, \ldots, q_N\}$ and $\mathcal{T} = \{t_1, t_2, \ldots, t_M\}$ respectively denotes a repository of medical records and their associated locally mined terminologies. The target of this section is to learn appropriate terminologies from the global terminology space $\mathcal{T}$ to annotate each medical record q in $\mathcal{Q}$. Among existing machine learning methods, graph-based learning achieves promising performance [28], [29]. In this work, we also explore the graph-based learning model to accomplish our terminology selection task, and expect this model is able to simultaneously considers various heterogeneous cues, including the medical record content analysis, terminology-sharing networks, the inter-expert as well as inter-terminology relationships. We will first introduce relationship identification and then we detail how to use our proposed model to link the underlying connected medical records. Next, we present the optimal solution for our learning model followed by the label bias estimation. Finally, we discuss the scalability of our method.

4.1 Relationship Identification

The inter-terminology and inter-expert relationships are not intuitively seen or implied from medical records. We thus call them as implicit relationships. This subsection aims to introduce how to discover these kinds of relationships.

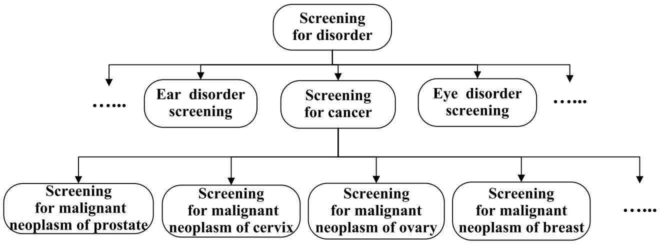
Screening for disorder
......
Ear disorder screening
Screening for cancer
Eye disorder screening
......
Screening for malignant neoplasm of prostate
Screening for malignant neoplasm of cervix
Screening for malignant neoplasm of ovary
Screening for malignant neoplasm of breast
......

Fig. 2. Partial illustration of SNOMED CT hierarchy for the terminology "screening for disorder".

4.1.1 Inter-Terminology Relationship

The medical terminologies in SNOMED CT are organized into acyclic taxonomic (is-a) hierarchies. For example, "viral pneumonia" is-a "infectious pneumonia" is-a "pneumonia" is-a "lung disease". Terminologies may also have multiple parents. For example, "infectious pneumonia" is also a child of "infectious disease". Fig. 2 shows part of the SNOMED CT hierarchy for the class of "screening for disorder". The well-defined ontology is able to semantically capture the inter-terminology hierarchical relationships. Given two terminologies t_i and t_j, their hierarchical relationship is quantitatively estimated as:

$$R_{ij} = \begin{cases} \frac{1}{2^p}, & \text{if ancestor-child relationships,} \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where p is the length of ancestor-child path between code t_i and t_j. And $\mathbf{R}$ is a matrix representing the weighted inter-terminology relationships.

The medical terminology hierarchy will enhance our scheme in two ways. First, it tackles the granularity mismatch problem, where the terminologies found in the medical records are very detailed and specific, while those in the query may be more general and high-level. This is achieved by rewarding the ancestral nodes with appropriate weights. Second, the hierarchical relationships boost the coding accuracy via filtering out the sibling terminologies. According to our observation, the sibling terminologies are rarely annotated for the same medical records, because they usually depict different body parts or emphasizes. For example, as shown in Fig. 2, the sibling nodes refer to non-overlapping disorders.

4.1.2 Inter-Expert Relationship

In this paper, the inter-expert relationships will be viewed stronger if the experts are professionals in the same or related specific medical areas. This is reflected by their historical data, i.e., the number of questions they have co-answered. Inspired by the Jaccard coefficient [30], the relationship between two experts u_i and u_j is calculated as

$$J(u_i, u_j) = \frac{|\mathcal{U}_i \cap \mathcal{U}_j|}{|\mathcal{U}_i \cup \mathcal{U}_j|}, \tag{9}$$

where $\mathcal{U}_i$ is the set of medical records that expert u_i have involved. The Jaccard coefficient is known to be useful to measure the similarity between two objects, which are represented by two unordered sets.

4.2 Probabilistic Hypergraph Construction

The graph-based learning models can be broadly categorized into simple graph-based and hypergraph-based approaches. They are both built on a graph where vertices are samples. While the simple graph only convey the pairwise relationship of vertices and overlooks the relations in higher orders, which are sensitive to the radius parameter used in similarity calculation [31]. As compared to simple graph, hypergraph contains the summarized local grouping information by allowing each hyperedge to connect more than two vertices simultaneously. Meanwhile hyperedge types and weights can be empirically set according to certain rules, and they can be heterogeneous to fuse comprehensive and diversified sources. Taken together, hypergraph-based learning partially fits our task of terminology selection via integrating multi-faceted information cues, except considering the inter-terminology hierarchical relationship. We extend this model to be applicable of our application.

A hypergraph $G(\mathcal{V}, \mathcal{E}, \mathbf{W})$ is composed of the vertex set $\mathcal{V}$, the hyperedge set $\mathcal{E}$, and the diagonal matrix of hyperedge weight $\mathbf{W}$. Here, $\mathcal{E}$ is a family of arbitrary subsets e of $\mathcal{V}$ such that $\cup_{e\in\mathcal{E}} = \mathcal{V}$, and each hyperedge e is assigned weight $W(e)$. A probabilistic hypergraph G can be represented by a $|\mathcal{V}| \times |\mathcal{E}|$ incidence matrix $\mathbf{H}$ with the following entries,

$$h(v_i, e_j) = \begin{cases} P(v_i, e_j), & \text{if } v_i \in e_j, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where $P(v_i, e_j)$ describes the probability that vertex v_i falls into the hyperedge e_j. Based on $\mathbf{H}$, the vertex degree of $v_i \in \mathcal{V}$ is estimated as,

$$d(v_i) = \sum_{e_j\in\mathcal{E}} W(e_j) h(v_i, e_j). \tag{11}$$

For a hyperedge $e_j \in \mathcal{E}$, its degree is defined as

$$\delta(e_j) = \sum_{v_i\in e_j} h(v_i, e_j). \tag{12}$$

We denote the vertex degrees and hyperedge degrees by $\mathbf{D}_v$ and $\mathbf{D}_e$, respectively.

In our work, the N medical records from $\mathcal{Q}$ are regarded as vertices and they are connected by three types of hyperedges. The first type takes each vertex as a centroid and forms a hyperedge by circling around its k-nearest neighbors based on medical record content similarities. This procedure was firstly adopted in [28]. The second type is based on terminology-sharing network. For each terminology, it groups all the medical records sharing the same terminology together. These two kinds of hyperedges highlight different semantic granularity. Suppose that the hypergraph vertices contain these two medical records: *"What are the signs of pregnancy in first weeks?"* and *"Is it safe to color my hair during pregnancy?"*. The terminology sharing network will link these two medical records together, because they both contain the medical concept "pregnancy", while the hyperedges that are based on semantic similarities among medical records may not group these two vertices together, since they belong to different health topics. Generally speaking, terminology sharing network is capable of capturing semantic relationships in sub-topic level via discrete binary method, and the content-based one is able to grasp the semantic relationship in high-level topic level via continuous quantization. They complement each other, instead of generating redundant information. The third kind actually takes the users' social behaviours into consideration by rounding up all the questions answered by closely associated experts. As a consequence, up to $N + M + U$ hyperedges are constructed in our hypergraph, where U is the number of involved experts. For each hyperedge, the likelihood of each constituent medical record belonging to its local group is defined according to its hyperedge type as follows:

$$P(v_i, e_j) = \begin{cases} 1 & \text{Inter-expert Relationships}, \\ K(\mathbf{q}_i, \mathbf{q}_j) & \text{Content Similarity}, \\ 1 & \text{Terminology-Sharing}, \end{cases} \tag{13}$$

where $K(\cdot,\cdot)$ is the Gaussian similarity function [32], which is a measure of similarity between two feature vectors, with considering the whole data distribution, mathematically stated as,

$$K(\mathbf{q}_i, \mathbf{q}_j) = exp\left(-\frac{||\mathbf{q}_i - \mathbf{q}_j||^2}{\sigma^2}\right), \tag{14}$$

where $\mathbf{q}_i$ is the feature vector for the ith medical record. The radius parameter σ is simply set as the median of the euclidean distances among all medical records. Then the initial weight for each hyperedge is,

$$W(e_j) = \sum_{v_i\in e_j} h(v_i, e_j). \tag{15}$$

The magnitude of the hyperedge weight indicates to what extent the vertices in a hyperedge belong to the same group [26].

Conventional hypergraph model has been widely used to solve many problems, such as community detection [33] and classification [34]. Take binary classification as an example, which is typically modulated as a regularization framework,

$$\arg\min_{\mathbf{f}} \Phi\ (\mathbf{f}) = \arg\min_{\mathbf{f}} \{\Omega(\mathbf{f}) + \lambda L(\mathbf{f})\}, \tag{16}$$

where vector $\mathbf{f}$ contains the relevance probabilities that we wish to learn. $\Omega(\mathbf{f})$ and $L(\mathbf{f})$ denote the regularizer on the hypergraph and empirical loss, respectively. The parameter λ is a regularization parameter to balance the empirical loss and the regularizer.

In this work, the medical terminology assignment task is regarded as a multi-label transductive learning problem. Inspired by Eq. (16), it is formulated as,

$$\arg\min_{\mathbf{F}} \Phi(\mathbf{F}) = \arg\min_{\mathbf{F}} \sum_{i=1}^{M} \{\Omega(\mathbf{f}_i) + \lambda L(\mathbf{f}_i)\}, \tag{17}$$

where M refers to the number of classes, i.e., the number of medical terminologies. Vector $\mathbf{f}_i$ is the ith column of $\mathbf{F}$, representing the relevance scores of each medical record to the ith code.

To explicitly incorporate the well-structured inter-terminology relationships, one more regularizer term need to be incorporated into Eq. (17):

$$\arg\min_{\mathbf{F}} \sum_{i=1}^{M} \left\{ \Omega(\mathbf{f}_i) + \lambda L(\mathbf{f}_i) + \mu \sum_{j=1}^{M} R_{ij} \|\mathbf{f}_i - \mathbf{f}_j\|^2 \right\}, \quad (18)$$

where R_{ij} is the inter-class relationship between class i and class j that is defined in Section 4.1.1. μ is a regularization constant to regulate the effect of the third term. Obviously, the ith row of $\mathbf{F}$ represents the relevance scores of all the terminologies to the medical record i. These relevance scores are reranked in descending order and the top c terminologeis are selected as the recommended results.

4.3 Global Learning Optimization

In Section 4.2, we have defined the hypergraph-based framework for the global terminology learning that contains three objectives. Here we aim to formulate each objective in details and derive a solution to this optimization problem. The philosophies to formulate these three objectives are as follows. The first objective should guarantee that the relevance probability function is continuous and smooth in semantic space. This means that the relevance probabilities of semantically similar medical records should be close to each other. The second objective is ensured by the empirical loss function, which forces the relevance probabilities to approach the initial roughly estimated relevance scores. These two implicit constraints are widely adopted in reranking-oriented approaches [35], [36]. The third objective encourages the values of medical records, which are connected by hierarchical structured terminologies, should be similar to each other.

Inspired by the normalized cost function of a simple graph [37], [38], $\Omega(\mathbf{f})$ is defined as

$$\frac{1}{2} \sum_{e \in \mathcal{E}} \sum_{u,v \in e} \frac{w(e) h(u,e) h(v,e)}{\delta(\mathbf{e})} \left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}} \right)^2. \quad (19)$$

By defining $\Theta = \mathbf{D}_v^{-\frac{1}{2}} \mathbf{H} \mathbf{W} \mathbf{D}_e^{-1} \mathbf{H}^T \mathbf{D}_v^{-\frac{1}{2}}$, we can further derive that

$$\Omega(\mathbf{f}) = \mathbf{f}^T (\mathbf{I} - \Theta) \mathbf{f}, \quad (20)$$

where $\mathbf{I}$ is an identity matrix. Let $\Delta = \mathbf{I} - \Theta$, which is a positive semi-definite matrix, the so-called hypergraph Laplacian [34], $\Omega(\mathbf{f})$ can be rewritten as,

$$\Omega(\mathbf{f}_i) = \mathbf{f}_i^T \triangle \mathbf{f}_i. \quad (21)$$

For the loss term, we introduce a new vector $\mathbf{y}$, which contains all the initially estimated relevance probabilities and define the loss term as a least square function as follows:

$$L(\mathbf{f}_i) = \|\mathbf{f}_i - \mathbf{y}_i\|^2 = \sum_{v \in \mathcal{V}} (f_i(v) - y_i(v))^2, \quad (22)$$

where $y_i(v)$ varies between $[0, 1]$ and is the initially estimated probability of sample v labeled with code i. Feeding Eqs. (21) and (22) into Eq. (18), we obtain $\Phi(\mathbf{F})$ as,

$$\sum_{i=1}^{M} \left\{ \mathbf{f}_i^T \triangle \mathbf{f}_i + \lambda \|\mathbf{f}_i - \mathbf{y}_i\|^2 + \mu \sum_{j=1}^{M} R_{ij} \|\mathbf{f}_i - \mathbf{f}_j\|^2 \right\}. \quad (23)$$

By minimizing $\Phi(\mathbf{F})$, it is able to satisfy our initial philosophies. We can solve $\mathbf{F}$ by intuitively and equivalently solving the following problem:

$$\arg\min_{\mathbf{F}} \{ \mathbf{F}^T \triangle \mathbf{F} + \lambda (\mathbf{F}^T \mathbf{F} - 2\mathbf{F}^T \mathbf{Y}) + \mu \mathbf{F} \mathbf{G} \mathbf{F}^T \}, \quad (24)$$

where $\mathbf{G}$ is the graph Laplacian of $\mathbf{R}$, $\mathbf{G} = \mathbf{V} - \mathbf{R}$, $\mathbf{V} = diag(\mathbf{V}_{jj})$, $\mathbf{V}_{jj} = \sum_i^M \mathbf{R}_{ij}$. By differentiating the above equation with respect to $\mathbf{F}$, we have

$$\mathbf{F} = \frac{\lambda}{1+\lambda} \left(\mathbf{I} + \frac{\mu \mathbf{G}}{1+\lambda} - \frac{\Theta}{1+\lambda} \right)^{-1} \mathbf{Y}. \quad (25)$$

4.4 Pseudo Label Estimation

As introduced in Section 4.3, our philosophy of the empirical loss term is to ensure the learnt relevance probabilities between terminologies and medical records are not far away from the initial roughly estimated relevance scores. In this section, we detail how to estimate the initial relevance scores.

$\mathbf{Y}_{N \times M}$ is a label biases matrix, where N and M respectively denotes the number of medical records and the number of terminologies. Y_{ij} stands for the initially estimated relevance between medical records i and terminology j. If terminology j is explicitly associated to medical record i identified by our local method, Y_{ij} is assigned to be 1. Otherwise, the well-known kernel density estimation approach [39], [40] is employed to estimate the relevance:

$$Y_{ij} = \frac{1}{|\chi_j|} \sum_{q_c \in \chi_j} K(\mathbf{q}_i, \mathbf{q}_c), \quad (26)$$

where χ_j denotes a set of medical records containing the terminology t_j. $K(.,.)$ is the Gaussian similarity function defined in Eq. (14). The above equation can be interpreted as follows: t_j and each of its associated medical record q_c in χ_j can respectively be viewed as a family and family members. Then the closeness of an unknown medical record to this family is estimated by averaging the soft voting from all family members.

4.5 Complexity Analysis

For the proposed graph-based global learning, the computational cost magnitude is analyzed as:

$$O\big(E^3 + 2NE^2 + 2EN^2 + N^3 + dN^2\big), \quad (27)$$

where d denotes the dimension of extracted features. N and E respectively represents the number of involved medical records and hyperedges. The hypergraph built on a medical records collection could be very large, and thus it is would be inefficient if we directly conduct the learning on such a hypergraph.

TABLE 1
Statistics of Our Data Collection

Unique Question #	Answer #	Duplicate Answer #	Unique Expert #
109,843	160,736	6,444	5,958

By partitioning the graph, the inference will be performed on a smaller scale, improving the efficiency to a great extent [41]. On the other hand, partitioning the entire hypergraph into multiple closely connected medical records capsules and learning the medical terminologies within each capsule will reduce the noise brought in by other irrelevant subgraphs.

Another approach to reduce the size of the hypergraph is by pre-clustering the medical records during the data collection stage into several sub-groups, and the hypergraph-based learning is only conducted within each cluster. Each cluster contains the semantically close medical records, which are inter-connected and most probably share the same terminologies. We adopt the pre-clustering technique in this study. With this preprocessing, the process can be completed within 1 second by a quad-core pentium processor of frequency: 3.4 GHz with 8 G memory. In this way, we are capable of providing instant recommendations of terminologies. And it can be employed to handle a much larger dataset.

4.6 Discussions

As detailed in Section 3.4, local mining approach suffers from three limitations: information loss, lower precision and over-widened terminology space problems. This subsection aims to discuss how global learning approach breaks these barriers.

The information loss is caused by some missing key concepts of the given medical record. They, however, are probably present in the semantically similar neighbours. Supposing t_j denotes the terminology corresponding to the missing key concept c_j in the given medical record q_i. According to our pseudo label estimation in Eq. (26), the relevance score between q_i and t_j will be initialized very high, because most of samples in χ_j are comparatively similar to q_i. The empirical loss function term in our model forces the learning relevance probabilities F_{ij} to approach Y_{ij}. In addition, the first regularizer term in our model also ensures that the relevance score between t_j and q_i should be close to the scores between t_j and the neighbours of q_i. Consequently, the global learning is able to discover the missing key concepts from underlying connected medical data and strongly link them to the given medical record.

The presence of irrelevant concepts results in the lower precision. Typically, these irrelevant medical concepts have grammatical meaning for communication between humans to help understand their intent, but they have less medical highlights and sparsely distribute in semantically similar data space. Even though Y_{ij} is initialized to be 1 and the empirical loss function attempts to make F_{ij} approaching 1, the first regularizer term will bring the relevance score down. Therefore, global learning is able to keep off these irrelevant concepts.

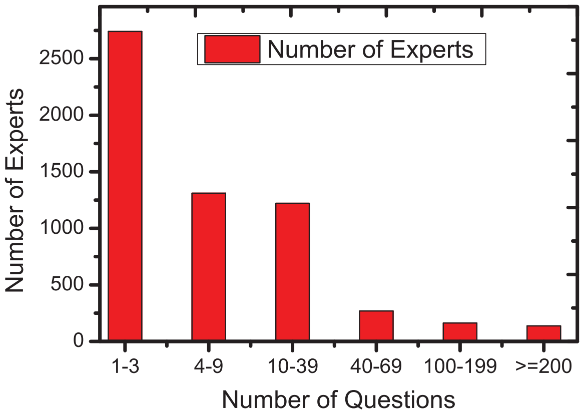

Fig. 3. The distribution of experts with respect to the number of questions they answered.

In this work, we address the over-widened space problem by merging all the locally assigned terminologies in our data collection to naturally form a corpus-aware terminology space, and utilize the proposed global learning approach to learn and propagate medical terminologies within this scope.

5 Experiments

In this section, we introduce the empirical evaluation of the proposed scheme. We first discuss the experimental settings, including the dataset and ground truth labeling. We then individually validate each component of our scheme. Finally, we comparatively evaluate the whole scheme.

5.1 Experimental Settings

We crawled more than 109 thousand medical records from HealthTap. Each medical record contains question, answers, and all the involved experts who answered the question before. Table 1 shows the statistics of our data collection. We can see that more than six thousand questions, though they may be lexically different, share the same answers. This shows that the vocabulary gap among users is very large. Figs. 3 and 4 show the analytic of our data collection, where around 54 percent of experts have replied to at least four questions and more than 33.2 percent of questions have at least two answers. These intersected structure is the basis of learning codes from neighbors.

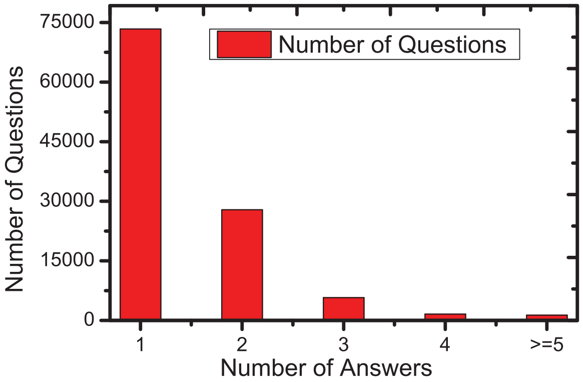

Fig. 4. The distribution of questions with respect to their received answers.

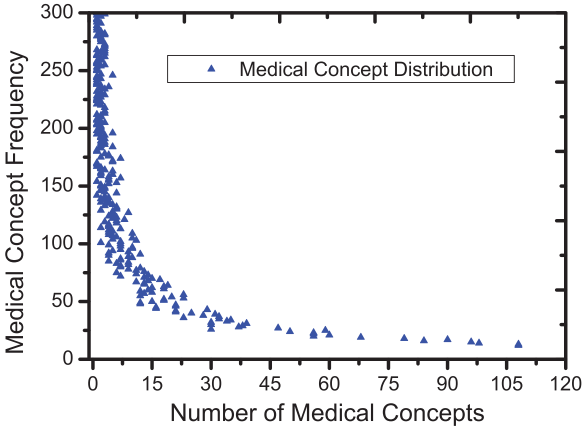

Fig. 5. The medical concept frequency distribution with respect to the number of distinct medical concepts.

The questions with only one answer or multiple answers but all from the same expert were eliminated, because they are isolated and unable to contribute much to the relationship investigation. Meanwhile, the experts who replied less than four questions, along with the associated questions, were also removed. These eliminations left us 36,473 questions. We have verified that removing non-active doctors does not result in severe loss of information. When we took a closer look at the dataset, we found that it is reasonable: the non-active doctors generally do not concern their online reputation and seem not carefully to answer the questions within their own expertise. Therefore, the non-active doctors bring in noises. On the other hand, if we consider these non-active doctors, the number of inter-expert based hyperedges will doubles in size, which decreases the efficiency. That is why we determined to remove them.

Unlike normal documents, these question samples are typically short, consisting of only one or two sentences. They thus do not provide sufficient word co-occurrences or shared contexts for effective similarity measure. It limits the accuracy obtained by the general learning methods. In our work, we incorporate the answers to supplement the short questions, which well compensate for the data sparseness issue.

For ground truth construction, we invited three professionals with master degrees majored in medicine programme. The labelers were trained with a short tutorial and a set of demonstrating examples. Although the ground truth labeling is subjective, a majority voting can alleviate the problem.

5.2 Local Mining Analysis

After data preprocessing, our pattern-based method extracted 13,845 unique noun phrases in total. The precision is ensured to be 100 percent since we defined the pattern according to the noun-phrase regular expression shown in Eq. (1).

As mentioned in Section 3.2, for each noun phrase, its specificity was estimated by comparing the term frequencies between two different corpora D_1 and D_2. D_1 is our medical-domain corpus and D_2 is a general English Gigaword data of Linguistic Data Consortium.[10] We detected 8,910 distinct medical concepts in total. Their frequency distribution is illustrated in Fig. 5, which roughly obeys the power law distribution. The medical concepts with occurrencing frequencies greater than five were selected to represent the content of each medical record. That is to say, each medical record will be represented by a 4,877-dimensional bag of medical concept histograms. However, it is unreasonable if we treat each dimension equally since we observed that the medical concepts with higher frequencies are usually more generic and less informative, such as the "women's health" and "pain". While medical concepts with rare occurrences are very specific and descriptive, such as "rotavirus infection" and "muscle paralysis". To this end, we adaptively weighted each medical concept in terms of its frequency, $r(c) = \frac{1}{\log(o(c)+1)}$, where $o(c)$ refers to the occurrence frequency of medical concept c. This formula stamps the generic medical concepts and rewards the specific medical concepts. Finally, each feature vector was normalized to have zero-mean and one-variance.

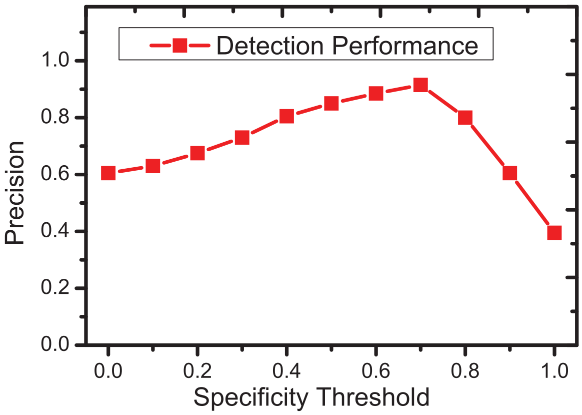

Fig. 6. Illustration of threshold selection for specificity.

To select the optimal threshold for specificity and validate the performance of medical concept detection, we randomly selected 400 noun phrases from the extracted noun phrase space, and equally split them into two subsets, one as threshold learning and the other as testing. Among these two sets of noun phrases, 116 and 121 were respectively voted as the medical concepts by our three annotators. Fig. 6 demonstrates the procedure of the specificity threshold selection. The peak value is reached when threshold arrives at 0.7. Table 2 illustrates the confusion matrix obtained by our proposed medical concept detection. We can see that our approach achieves fairly good performance, i.e., 91.5 percent. The misclassified results mainly come from two parts. First, some concepts, such as "mosquito allergy" and "kidney stone removal" sparsely distribute in our medical dataset and they can not be detected as medical concepts even they are. This is because their domain-relevances are

TABLE 2
The Confusion Matrix of Medical Concept Detection Results

Prediction Class	Medical Concepts	Non-medical Concepts
Medical Concepts	110	6
Non-medical Concepts	11	73

The prediction accuracy is 91.5 percent.

[10] http://www.ldc.upenn.edu/.

TABLE 3
The 10 Representative Medical Concepts and Their Corresponding Terminologies After Normalization

Medical Concepts	Normalized Terminologies
Birth control	Contraception
Blood loss	Hemorrhage
Breast cancer	Malignant tumor of breast
Condom	Uses contraceptive sheath
Home pregnancy test	Pregnancy test finding
Late menses	Menstrual period late
Ovarian cyst	Cyst of ovary
Period pain	Dysmenorrhea
Sex	Finding of sexual intercourse
Spontaneous abortion	Spontaneous abortion

estimated very low by our comparative measure. On the other hand, some non medical concepts, such as "excise" and "walkers", have been frequently mentioned by experts, which make them classified as medical concepts.

During medical concept normalization, some interesting phenomenons were observed:

1) Not all the detected medical concepts can be mapped to one entry in SNOMED CT. For example, some experts misspelt "menses" as "mense", while "mense" is not searchable.
2) Multiple medical concepts may be converted into the same terminology. For example, both "chloasma gravidarum" and "pregnancy mask" refer to "melasma gravidarum". This further verifies that the vocabulary varies widely among content generators.
3) Less than 15 percent of medical concepts are the same with their normalized terminologies. This implies that the usage of authenticated terminologies is very sparse in social writings and it is highly desired to normalize them.

Our approach is able to pair each medical concept and its corresponding terminology. We then randomly selected 100 medical concept-terminology pairs to validate our method. Three annotators then voted each pair as "correct" or "incorrect". Our proposed normalization approach achieved up to 82 percent accuracy. The coarse-fine granularity is a leading cause of errors. Take "menstrual cycle" for an example. Our approach was unable to find the same granularity of terminology related to this medical concept. Instead, numerous specific-level terminologies, such as "long menstrual cycle", "short menstrual cycle", and "abnormal menstrual cycle" were derived. These specific-level terminologies, however, sometimes were viewed as incorrect by annotators to represent the original medical concepts. Table 3 displays ten representative medical concept-terminology pairs.

5.3 Graph-Based Global Learning Analysis

To demonstrate the effectiveness of our global learning approach, we compare it against the following state-of-art learning approaches:

- *PRFeedback*. Pseudo-Relevance Feedback [42]. For each terminology, a SVM classifier was trained to estimate the relevance score between this terminology and each medical record. The training data was generated based on the assumption that the top-ranked samples are more relevant than the lower-ranked results in general. The initial medical record ranking list with respect to this terminology was generated based on Eq. (26). This method requires to train M classifiers. (Baseline 1)
- *RWReranking*. Random Walk based Reranking [29]. For each given medical record q, a simple graph was first constructed. The stationary probability output the relevance scores between q and any other medical records involved in the graph. The terminologies associated with the top 50 medical records were selected as candidates, and their relevance score to q was estimated as, $P(q,t) = \frac{\sum_{q_t \in \mathcal{Q}_t} p(q_t,q)}{|\mathcal{Q}_t|}$, where $\mathcal{Q}_t$ denotes the medical record set containing terminology t in the top 50 medical records.(Baseline 2)
- *CHLearning*. Conventional hypergraph learning [28]. The inter-terminology hierarchical relationships were not considered at all. Specifically, the results were inferred from Eq. (17) instead of Eq. (18). (Baseline 3)
- *GGLearning*. Our global learning approach.

To fairly evaluate our unsupervised learning approach, other supervised graph-based learning methods were not listed here. For each method mentioned above, the involved parameters were carefully tuned, and the parameters with the best performances were used to report the final comparison results. For example, the two parameters λ and μ were obtained with grid search in flexible step size to achieve optimal coding performance in terms of NDCG@5, which respectively are 9 and 1. Meanwhile, these four methods were all based on local mining results, i.e., each medical record has been locally coded.

The ground truth was created by a manual labelling procedure. To be specific, we randomly selected 100 medical records and for each one we generated four different ranking lists of terminologies by the above methods. Three annotators were asked to label the top 20 terminologies to be very relevant (score 2), relevant (score 1) or irrelevant (score 0), with respect to the given medical record. We performed a voting to establish the final relevance level of each terminology.

The inter-rater reliability of the labeling task was analyzed with the Kappa method [43]. The Kappa metric is a chance corrected statistic to quantitatively measure the degree of inter-rater agreement. It can be interpreted as expressing the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters made their ratings completely randomly. Kappa result ranges from 0 to 1. Kappa value more than 0.7 typically indicates that agreement is strong. The results demonstrated there were sufficient inter-rater agreements. In our work, there are 2,000 cases, three categories, and three raters. The fixed-marginal kappa and free-marginal kappa values are respectively 0.761 and 0.8.

NDCG@n [39] was adopted as our metric to measure the performance of various learning approaches. The comparison results are illustrated in Fig. 7. It is observed that **CHLearning** and **GGLearning** are consistently and remarkably better than the other two baselines across various evaluating depth of NDCG. One possible reason is the

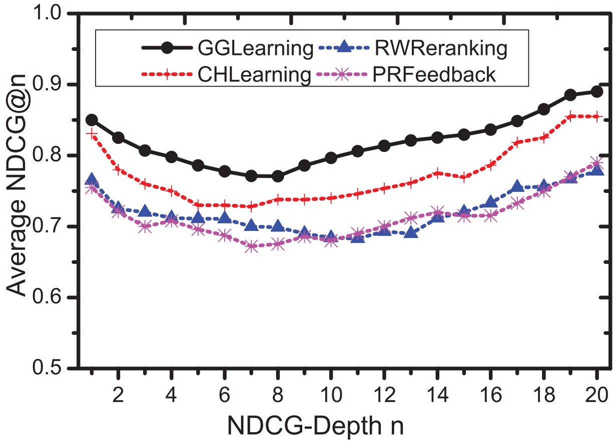

Fig. 7. Performance comparison of global learning-based medical terminology assignment.

unreliable initial ranking list resulted by the rough estimation. The second reason is probably because the hypergraph-based learning is able to capture the high-order relationships among medical records, i.e., the summarized local grouping information, in contrast to simple pairwise relationships characterized by other two approaches. In addition, hypergraph-based learning is able to integrate heterogeneous information cues, such as terminology-sharing network and inter-expert relationships, rather than the pure lexical property based similarity. From this figure, we can also observe that our proposed method consistently outperforms **CHLearning**. This is because the inter-terminology hierarchical relationships mined from external knowledge are leveraged by our proposed global learning approach to enhance the coding performance.

5.4 Medical Terminology Assignment

It is well known that for the annotation task, precision is usually more important than recall. We thus adopted two metrics that are able to characterize precisions from different aspects. The first one is average $S@K$ over all testing medical records, which measures the probability of finding a relevant terminology among the top K recommended ones. To be specific, for each testing medical record, $S@K$ is assigned to be 1 if a relevant terminology is positioned in the top K and 0 otherwise. The second one is average $P@K$ that stands for the proportion of recommended terminologies that are relevant. $P@K$ is defined as, $P@K = \frac{|\mathcal{C} \cap \mathcal{R}|}{|\mathcal{C}|}$ where $\mathcal{C}$ is a set of the top K terminologies, and $\mathcal{R}$ is the manually labeled positive ones. The ground truth construction is analogous to Section 5.3. The slight differences are that the invited annotators were required to label only top five suggested terminologies for each medical record, and they were labeled either as "positive" or "negative".

We comparatively evaluate our proposed medical terminology assignment scheme with other competitive coding approaches:

- *TagCollective*. Tag Recommendation based on Collective Knowledge [30]. This approach is a statistical and data-driven method. To be specific, given a medical record with locally mined terminologies, an ordered list of m terminology candidates were derived for each of the locally mined terminologies based on the their co-occurrences. The lists of candidate terminologies were then used as input for terminology aggregation and ranking, which ultimately produces the ranked list of n recommended terminologies. For the aggregation, we employed the vote-based strategy as introduced in [30].
- *TagAssist*. This method was first introduced by [44]. It annotates a medical record by generating the search queries from the given medical record, searching a collection of medical records using those queries, and selecting suitable terminologies from the retrieved medical records. For the matching, we employed the method in [42].
- *LocalMining*. Our local mining approach.
- *Local+Global*. Our proposed joint local-global scheme for medical terminology assignment.

TABLE 4
The Comparative Evaluation Results of Medical Terminology Assignment in Terms of $S@K$

Apporaches Metrics	S@1	S@2	S@3	S@4	S@5
TagCollective	76.0%	87.0%	95.0%	98.0%	99.0%
TagAssist	78.0%	88.0%	96.0%	98.0%	100.0%
LocalMining	72.0%	84.0%	91.0%	95.0%	96.0%
Local+Global	**83.0%**	**92.0%**	**98.0%**	**100.0%**	**100.0%**

The comparison results in terms of $S@K$ and $p@K$ are respectively displayed in Tables 4 and 5. We can see that the local mining approach achieves the worst performance, since irrelevant concepts may be mapped to terminologies because of their presence in the medical records. It is also observed that our proposed medical terminology assignment scheme significantly outperforms the others. Our approach achieves up to 100 percent at $S@4$, which suggests that at least one recommended medical terminology can be ensured to be relevant in the top four terminologies. On the other hand, $p@5$ larger than 0.76 means that approximate four terminologies on average are relevant in the top five recommended ones. This demonstrates the effectiveness of our graph-based global learning component. **TagCollective** only statistically considers the co-occurrences among terminologies, which completely ignores the lexical property-based similarities among medical records, let alone other complex characteristics and higher order analytic. Even though the **TagAssist** takes pairwise similarities into consideration, it does not consider the hierarchical relationships among terminologies or the social connections among experts. Meanwhile, we can see that the performance of

TABLE 5
The Comparative Evaluation Results of Medical Terminology Assignment in Terms of $P@K$

Apporaches Metrics	P@1	P@2	P@3	P@4	P@5
TagCollective	76.0%	75.5%	74.3%	72.8%	71.0%
TagAssist	78.0%	77.0%	75.6%	74.3%	72.8%
LocalMining	72.0%	72.1%	69.7%	68.3%	66.6%
Local+Global	**83.0%**	**81.5%**	**80.3%**	**78.8%**	**76.4%**

TABLE 6
Comparative Illustration of the Representative Question Samples with Locally Mined Terminologies and Locally+Globally Recommended Terminologies

Medical Records	Locally Mined Terminologies	Local Mining + Global Learning
Is it safe to color my hair during pregnancy ?	hair structure, dyed hair, feeling safe, patient currently pregnant, first trimester pregnancy...	hair structure, patient currently pregnant, coal tar allergy, hair color change, disorder of endocrine system...
What are the risks getting pregnant and giving birth later in life ?	finding of at risk, cesarean section, birth, patient currently pregnant, finding of life event...	finding related to risk factor in pregnancy, birth, advanced maternal age gravida, diabetes mellitus during pregnancy patient currently pregnant...
If I get an infection caused by gum disease, can that be transferred to my fetus ?	infectious disease, gingival disease, entire fetus, inflammation, periodontal disease...	infectious disease, prematurity of fetus, gingival disease, periodontal disease low birth weight infant...

Answers are not displayed due to limited space.

TagAssist is a bit better than **TagCollective**, which reflects that the lexical property-based similarities is more reliable than terminologies co-occurrence cues.

We also performed an ANalysis Of VAriance (ANOVA) test with single factor only. It was conducted over average $P@K$ between our scheme and each of the baselines. The F and p values between **Local+Global** and **LocalMining** are respectively 43.54 and 0.00017. These two values between **Local+Global** and **TagAssist** are 9.26 and 0.016, respectively. And they are respectively 17.40 and 0.0031 between **Local+Global** and **TagCollective**. It can be seen that all p values are much smaller than 0.05, which indicates that the improvements are statistically significant.

The following three figures respectively illustrate the results of statistical analysis. It can be seen that all p values are much smaller than 0.05, which indicates that the improvements are statistically significant.

Table 6 comparatively illustrates the representative medical record samples with locally minded terminologies and locally+globally recommended ones. From this table, we can see that the locally mined terminologies may be irrelevant to the medical records, such as “feeling safe” to the first example. This is caused by the appearance of “safe” in the medical record. Also we can observe that some key terminologies missed by local mining approach, such as “advanced maternal age gravida” to the second example. This may be resulted in by the absence of some key medical concepts of the original medical records. Intuitively, the terminologies are more comprehensive and reliable after enhancement with global learning. This is because it is able to learn terminologies from neighbors that complements the missing information and keeps off the irrelevant information.

6 Conclusions and Future Work

This paper presents a medical terminology assignment scheme to bridge the vocabulary gap between health seekers and healthcare knowledge. The scheme comprises of two components, local mining and global learning. The former establishes a tri-stage framework to locally code each medical record. However, the local mining approach may suffer from information loss and low precision, which are caused by the absence of key medical concepts and the presence of the irrelevant medical concepts. This motivates us to propose a global learning approach to compensate for the insufficiency of local coding approach. The second component collaboratively learns and propagates terminologies among underlying connected medical records. It enables the integration of heterogeneous information. Extensive evaluations on a real-world dataset demonstrate that our scheme is able to produce promising performance as compared to the prevailing coding methods. More importantly, the whole process of our approach is unsupervised and holds potential to handle large-scale data.

In the future, we will investigate how to flexibly organize the unstructured medical content into user needs-aware ontology by leveraging the recommended medical terminologies.

References

[1] L. Nie, M. Akbari, T. Li, and T.-S. Chua, “A joint local-global approach for medical terminology assignment,” in *Proc. Int. ACM SIGIR Conf.*, 2014.

[2] L. Nie, T. Li, M. Akbari, and T.-S. Chua, “Wenzher: Comprehensive vertical search for healthcare domain,” in *Proc. Int. ACM SIGIR Conf.*, 2014, pp. 1245–1246.

[3] AHIMA e-HIM Work Group on Computer-Assisted Coding, “Delving into computer-assisted coding,” *J. AHIMA*, vol. 75, pp. 48A–48H, 2004.

[4] G. Leroy and H. Chen, “Meeting medical terminology needs-the ontology-enhanced medical concept mapper,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 5, no. 4, pp. 261–270, Dec. 2001.

[5] G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt, “Exploiting medical hierarchies for concept-based information retrieval,” in *Proc. Australasian Document Comput. Symp.*, 2012, pp. 111–114.

[6] E. J. M. Lauría and A. D. March, “Combining Bayesian text classification and shrinkage to automate healthcare coding: A data quality analysis,” *J. Data Inf. Quart.*, vol. 2, no. 3, p. 13, 2011.

[7] L. Yves A., S. Lyudmila, and F. Carol, “Automating ICD-9-cm encoding using medical language processing: A feasibility study,” in *Proc. AMIA Annu. Symp.*, 2000, p. 1072.

[8] C. Dozier, R. Kondadadi, K. Al-Kofahi, M. Chaudhary, and X. Guo, “Fast tagging of medical terms in legal text,” in *Proc. Int. Conf. Artif. Intell. Law*, 2007, pp. 253–260.

[9] L. V. Lita, S. Yu, S. Niculescu, and J. Bi, “Large scale diagnostic code classification for medical patient records,” in *Proc. Conf. Artif. Intell. Med.*, 1995.

[10] L. S. Larkey and W. B. Croft, “Automatic assignment of icd9 codes to discharge summaries,” PhD dissertation, Dept. Comput. Sci., Univ. Massachusetts at Amherst, Amherst, MA, USA, 1995.

[11] H. Suominen, F. Ginter, S. Pyysalo, A. Airola, T. Pahikkala, S. Salanter, and T. Salakoski, “Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: A method description,” in *Proc. ICML Workshop Mach. Learn. Health-Care Appl.*, 2008.

[12] J. Patrick, Y. Wang, and P. Budd, "An automated system for conversion of clinical notes into snomed clinical terminology," in *Proc. 5th Australasian Symp. ACSW Frontiers*, 2007, pp. 219–226.

[13] W. R. Hersh and H. David, "Information retrieval in medicine: The saphire experience," *J. Amer. Soc. Inf. Sci.*, vol. 46, no. 10, pp. 743–747, 1995.

[14] Q. Zhou, W. W. Chu, C. Morioka, G. H. Leazer, and H. Kangarloo, "Indexfinder: A method of extracting key concepts from clinical texts for indexing," in *Proc. AMIA Annu. Symp.*, 2003, pp. 763–767.

[15] Y. Wang and J. Patrick, "Mapping clinical notes to medical terminology at point of care," in *Proc. Workshop Current Trends Biomed. Natural Lang. Process.*, 2008, pp. 102–103.

[16] S. Hina, E. Atwell, and O. Johnson, "Semantic tagging of medical narratives with top level concepts from SNOMED CT healthcare data standard," *Int. J. Intell. Comput. Res.*, vol. 2, pp. 204–210, 2010.

[17] H. Stenzhorn, E. Pacheco, P. Nohama, and S. Schulz, "Automatic mapping of clinical documentation to SNOMED CT," *Studies Health Technol. Inform.*, vol. 158, pp. 228–232, 2009.

[18] K. Crammer, M. Dredze, K. Ganchev, P. P. Talukdar, and S. Carroll, "Automatic code assignment to medical text," in *Proc. Workshop Biol., Translational, Clinical Lang. Process.*, 2007, pp. 129–136.

[19] L. R. S. de Lima, A. H. F. Laender, and B. A. Ribeiro-Neto, "A hierarchical approach to the automatic categorization of medical documents," in *Proc. Int. Conf. Inf. Knowl. Manag.*, 1998, pp. 132–139.

[20] Y. Yan, G. Fung, J. G. Dy, and R. Rosales, "Medical coding classification by leveraging inter-code relationships," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2012, pp. 193–202.

[21] S. V. Pakhomov, J. D. Buntrock, and C. G. Chute, "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques," *J. Amer. Med. Inf. Assoc.*, vol. 13, no. 5, pp. 516–525, 2006.

[22] Y. Chen, Z. Chenqing, and K.-Y. Su, "A joint model to identify and align bilingual named entities," *Comput. Linguistics*, vol. 39, no. 2, pp. 229–266, 2013.

[23] P. Bansal, S. Bertels, T. Ewart, P. MacConnachie, and O. James, "Bridging the research–practice gap," *Acad. Manag. Perspectives*, vol. 26, pp. 73–91, 2012.

[24] N. Chu, Y. Choi, J. Wei, and A. Cheok, "Games bridging cultural communications," in *Proc. IEEE Global Conf. Consumer Electron.*, 2012, pp. 329–332.

[25] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303–319, Apr. 2011.

[26] M.-Y. Kim and R. Goebel, "Detection and normalization of medical terms using domain-specific term frequency and adaptive ranking," in *Proc. IEEE Int. Conf. Inf. Technol. Appl. Biomed.*, 2010, pp. 1–5.

[27] R. L. Cilibrasi and P. M. B. Vitanyi, "The google similarity distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.

[28] Y. Huang, Q. Liu, S. Zhang, and D. Metaxas, "Image retrieval via probabilistic hypergraph ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3376–3383.

[29] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking through random walk over document-level context graph," in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 971–980.

[30] B. Sigurbjörnsson and R. van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 327–336.

[31] J. Yu, D. Tao, and M. Wang, "Adaptive hypergraph learning and its application in image classification," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3262–3272, Jul. 2012.

[32] H. Yang, L. Henry J., K. Dan, and C. Russell J., "Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS specialist lexicon," *J. Amer. Med. Informat. Assoc.*, vol. 12, no. 3, pp. 275–285, 2005.

[33] Y.-R. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, and A. Kelliher, "Metafac: Community discovery via relational hypergraph factorization," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2009, pp. 527–536.

[34] D. Zhou, J. Huang, and B. Schlkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1601–1608.

[35] L. Nie, Y.-L. Zhao, X. Wang, J. Shen, and T.-S. Chua, "Learning to recommend descriptive tags for questions in social forums," *ACM Trans. Inf. Syst.*, vol. 32, no. 1, p. 5, 2014.

[36] L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua, "Beyond text QA: Multimedia answer generation by harvesting web information," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 426–441, Feb. 2013.

[37] L. Nie, M. Wang, Z.-J. Zha, and T.-S. Chua, "Oracle in image search: A content-based approach to performance prediction," *ACM Trans. Inf. Syst.*, vol. 30, no. 2, p. 13, 2012.

[38] L. Nie, M. Wang, Z.-J. Zha, G. Li, and T.-S. Chua, "Multimedia answering: Enriching text QA with media information," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 695–704.

[39] L. Nie, S. Yan, M. Wang, R. Hong, and T.-S. Chua, "Harvesting visual concepts for image search with complex queries," in *Proc. Int. Conf. Multimedia*, 2012, pp. 59–68.

[40] Y.-L. Zhao, L. Nie, X. Wang, and T.-S. Chua, "Personalized recommendations of locally interesting venues to tourists via cross region community matching," *ACM Trans. Intell. Syst. Technol.*, vol. 9, p. 1, 2013.

[41] L. Li and T. Li, "News recommendation via hypergraph learning: Encapsulation of user behavior and news content," in *Proc. ACM 6th Int. Conf. Web Search Data Mining*, 2013, pp. 305–314.

[42] Y. Rong, H. Er, and J. Rong, "Multimedia search with pseudo-relevance feedback," in *Proc. Int. Conf. Image Video*, 2003, pp. 238–247.

[43] M. Warrens, "Inequalities between multi-rater kappas," *Adv. Data Anal. Classification*, vol. 4, no. 4, pp. 271–286, 2010.

[44] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua, "Bayesian video search reranking," in *Proc. ACM Int. Conf. Multimedia*, 2008, pp. 131–140.

Liqiang Nie received the BE degree from the Xi'an Jiaotong University of China, Xi'an, in 2009, and the PhD degree from the National University of Singapore, in 2013. He is currently a research fellow in the School of Computing, National University of Singapore. His research interests include information retrieval and healthcare analytics. Various parts of his work have been published in top forums including *ACM SIGIR*, *ACM MM*, *TOIS*, and *TMM*. He has been served as reviewers for various journals and conferences. He is a member of the IEEE.

Yi-Liang Zhao received the BEng degree in the School of Computer Engineering from Nanyang Technological University with first class honours. He is working toward the PhD degree from the School of Computing of National University of Singapore. He is an assistant professor in the Digipen Institute of Technology Singapore. He received the Research Achievement Award by the School of Computing of NUS in 2013 and Best Paper Award in the international conference on multimedia modeling in 2011. His currently research interests include social media analysis, community learning, and recommender systems.

Mohammad Akbari received the BSc degree in computer engineering from Teacher Training University, Tehran, Iran, in 2003, and the MSc degree in computer science and intelligent system from Amirkabir University of Technology, Tehran, Iran, in 2005. He is working toward the PhD degree at Graduate School for Integrative Sciences and Engineering, National University of Singapore. His current research interests include multimedia information retrieval, computer vision and machine learning for large scale content understanding.

Jialie Shen received the PhD degree in computer science from the University of New South Wales, Australia, in the area of large-scale media retrieval and database access methods. He is an assistant professor in the School of Information Systems, Singapore Management University, Singapore. His main research interests include information retrieval in the textual and multimedia domain, economic-aware media analysis and multimedia systems. His recent work has been published or is forthcoming in various leading journals and international conferences. He is a member of the IEEE.

Tat-Seng Chua is the KITHCT chair professor at the School of Computing, National University of Singapore (NUS). His main research interests includes multimedia information retrieval, multimedia question-answering, and the analysis and structuring of user-generated contents. He works on several multi-million-dollar projects: interactive media search, local contextual search, and real-time live media search. He has organized and served as program committee member of numerous international conferences in the areas of computer graphics, multimedia, and text processing. He is a member of the IEEE.