Akbari *et al. EURASIP Journal on Bioinformatics and Systems Biology* (2016) 2016:18
DOI 10.1186/s13637-016-0053-x

EURASIP Journal on Bioinformatics and Systems Biology

RESEARCH Open Access

Towards organizing health knowledge on community-based health services

Mohammad Akbari[1,2*], Xia Hu[3], Liqiang Nie[2] and Tat-Seng Chua[1,2]

Abstract

Online community-based health services accumulate a huge amount of unstructured health question answering (QA) records at a continuously increasing pace. The ability to organize these health QA records has been found to be effective for data access. The existing approaches for organizing information are often not applicable to health domain due to its domain nature as characterized by complex relation among entities, large vocabulary gap, and heterogeneity of users. To tackle these challenges, we propose a top-down organization scheme, which can automatically assign the unstructured health-related records into a hierarchy with prior domain knowledge. Besides automatic hierarchy prototype generation, it also enables each data instance to be associated with multiple leaf nodes and profiles each node with terminologies. Based on this scheme, we design a hierarchy-based health information retrieval system. Experiments on a real-world dataset demonstrate the effectiveness of our scheme in organizing health QA into a topic hierarchy and retrieving health QA records from the topic hierarchy.

Keywords: Consumer health information, Community question-answering, Information organization, Information retrieval

1 Introduction

The emergence of online health information needs has given rise to the establishment of online health services. Broadly speaking, current online health services can be divided into two categories. The first is the professional health provider released sources, such as Yahoo! Health[1] and WebMD[2]. These sources provide trustworthy and formally-written health information. They are usually well-structured in terms of health topics. The second category is the community-based health services (CHSs), such as HealthTap[3] and HaoDF[4]. These services allow health seekers to freely post health-oriented questions, and encourage doctors to provide quality answers. Compared to the former sources, CHSs have some intrinsic properties. First, they are crowdsourcing data that are continually growing at a fast pace, and it is thus not practical to organize them manually. Second, they are unstructured and unlabeled in terms of topics, which greatly hinder their retrieval and browsing by user. Third, health seekers and doctors with diverse backgrounds tend to present the same concepts in colloquial style, which leads to a wide vocabulary gap. Together, these pose big challenges for data access and navigation. Recent efforts [1] indicate that organizing the community-contributed data into a hierarchical structure may enhance coarse-grained browsing and fined-grained search.

Several practical systems and research efforts have been dedicated to organizing community-contributed data [1, 2]. Most of these efforts, however, suffer from the following limitations. First, they typically utilized predefined taxonomies in the form of tree structures and expect users or computers to assign data instances into these taxonomies based upon their understanding. However, the available taxonomies in health domain are usually too shallow with broad categorizations. For example, Yahoo! Answer[5] partitions health data into only nine main categories which are too general to summarize the diverse health information. Some popular topics such as "pregnancy" cannot be directly browsed here, because they do not fall under the predefined fixed category structure. Besides, these fixed taxonomies usually face the problems of being too centralized, conservative, and ambiguous [3]. Moreover, manual assignment by health seeker is probably

*Correspondence: akbari@u.nus.edu
[1] School for Integrative Sciences and Engineering, NUS, Singapore, Singapore
[2] School of Computing, NUS, Singapore, Singapore
Full list of author information is available at the end of the article

not applicable since they do not sufficiently understand their health problems. Second, existing efforts enable each data instance to be assigned into only one leaf node of the hierarchy. However, the health records are usually more verbose and complex, and probably convey multiple concerns. They hence should be assigned into more topic-level leaf nodes. Third, topic hierarchy construction approaches in general domain often annotate each node of the hierarchy with frequent occurrence terms or concepts. However, in vertical domain hierarchy construction, such as health domain, labeling nodes with standard terminologies is preferable, since it facilitates data reusability and exchange. Fourth, the existing efforts are unable to adaptively build the skeleton hierarchy. Specifically, the number of children for each given parent node and the number of layers in the entire hierarchy are either extracted from existing external structures or predefined by the so-called domain experts. They are often biased towards specific context or personal perspectives [4].

To overcome these limitations, we propose a top-down scheme that can organize the unstructured health records into a structured hierarchical tree. First, nodes in higher layers of the tree represent abstract topics. These nodes usually do not have clear definition and are thus difficult to be extracted automatically. On the other hand, even though the existing health-related taxonomies are very general, they still capture the high-level structures of the health domain well. We naturally leverage such prior domain knowledge to construct the higher layers of our hierarchy. Second, we propose an expanding approach to perform overlapping partitioning of each node to generate its children. Starting from the higher layer node, we try to obtain a hierarchy of its children. However, without termination criteria, the generated tree will be very huge in which each leaf node may contain only one health record. To address this problem, we propose a shrinkage approach to monitor and infer whether the node is specific enough before expansion. Following the breadth-first tree traversal trajectory, we alternatively employ expansion and shrinkage approaches to inspect each node and generate a proper hierarchy. In addition, all involved nodes are profiled with terminologies selected from the Unified Medical Language System (UMLS) Metathesaurus[6].

Based on our proposed organization scheme, we develop a hierarchy-based health information retrieval system. Health information search has attracted intensive attentions from industry and academia [5–9]. The effectiveness and efficiency of these efforts, however, are limited due to the inconsistent terms used in health domain and the need for exhaustive search in the entire data corpus. Our application adopts the topic-based matching and performs intelligent pruning of irrelevant branches of the generated hierarchy, and it can boost search performance significantly.

The contributions of our work are threefold:

- To the best of our knowledge, this is the first work on automatic organization of community-contributed health data.
- With prior domain knowledge, we propose a top-down organization scheme where skeleton hierarchy is automatically determined, multiple relations are enabled, and nodes are profiled with terminologies.
- We propose a hierarchy-based health information retrieval system. Extensive evaluations demonstrate its promising performance.

The remainder of this paper is organized as follows. Sections 2 and 3, respectively, detail our organization scheme and our hierarchy-based health information retrieval. Section 4 introduces the representation of QA records and similarity measures used. Experimental results and analysis are presented in Section 5. Section 6 reviews the related work, followed by our conclusion and future work in Section 7.

2 Top-down organization scheme

This paper targets at generating a rooted, directed, and profiled tree H from a given data corpus that contains n health-related question answering (QA) records $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$. Each node $\mathcal{V}$ in H is a subset of $\mathcal{D}$, representing a latent topic of semantically similar records. Notably, the root node $\mathcal{V}_0$ involves all the records in $\mathcal{D}$. The child nodes loosely partition their parent nodes, where overlapping is allowed. Figure 1 representatively shows the loose partitioning of the given parent node. From this figure, it can be seen that one health QA record can be assigned into two or more sibling nodes.

2.1 Incorporation of domain knowledge

As aforementioned, the current health-related taxonomies are usually very general and shallow. For example, the taxonomies provided by WebMD and Yahoo! Health are almost flat. They typically capture the high-level categorizations and structures of health domain. They are user-oriented categories which model human expectation of abstract categories in health records. On the other hand, automatic extraction of high-level categories of a given corpus is non-trivial, since there are overlaps and inter-correlation between topics especially in health domain. Take the categories of “mental health” and “women’s health” as an example; they are partially overlapped rather than being mutually exclusive and complementary. Regarding the aforementioned discussion, we employed such a domain knowledge to guide the construction of topic hierarchy and ensure that the generated structure is human readable and interpretable. While different kinds of domain knowledge may be available, in this

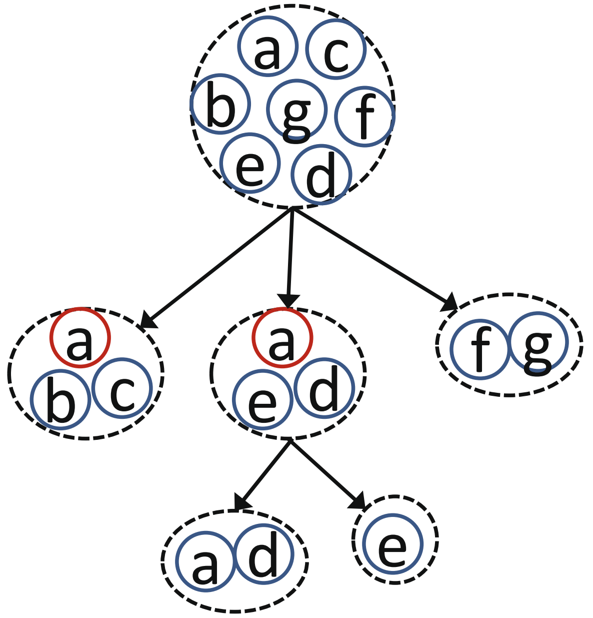

Fig. 1 Illustration of loose partition of parent node. *Dashed* and *solid circles* stand for nodes and health QA records, respectively

paper, we assume that prior domain knowledge is available as a predefined hierarchy structure. The predefined hierarchy structure may include several layers of nodes labeled with keywords. To facilitate the formalization of our hierarchy generation, in this paper, we utilized a one-level tree structure which includes several child nodes following the root node.

We utilize the categorization of healthexchange[7] as our initial first layer following the root node. Having a predefined hierarchy structure, we construct a set of classifiers to categorize health QA records in the root node into these categories. To accomplish this task, we first extract a set of exemplar QA pairs to represent the semantic context of each category. To do so, we employ each category's name as a query and obtain the top 100 relevant QA pairs from HealthTap. To form negative samples, we randomly select 100 negative samples from the other categories. We then trained a SVM classifier using the samples for each category.

2.2 Expanding approach

Through incorporating domain knowledge, we have partitioned the root node into a list of high-level categories which correspond to user expectation of knowledge structure. Each category is viewed as a node in the first layer. This subsection details the expanding approach to further generate a fine-grained hierarchy.

According to our definition, each node $\mathcal{V}$ in the target hierarchy H is a set of health QA records. We assume that this collection of health QA records can be explained by a set of unobserved abstract groups, and each group contains a small set of semantically similar health QA records talking about the same health topic. We then naturally shift our expanding task into topic modeling problem. The latent Dirichlet allocation (LDA) model [10] is utilized here, which is a generative model for discovering the unobserved abstract groups that occur in a data collection.

The main challenge in the expanding phase is to determine the proper number of children for each given node. Each child node should represent one aspect of the parent node, and complement to its siblings instead of mutually overlapping. Our proposed expansion approach selects the number of children via a tuning procedure. This procedure seeks for the children number that minimizes the LDA model's perplexity [10] on a held-out testing data set. It is formulated on a hold-out set with m health QA records as

$$\text{perplexity} = \exp\left\{\frac{\sum_{i=1}^{m}\log p(d_i)}{\sum_{i=1}^{m} l_i}\right\}, \tag{1}$$

where l_i is the length of health QA record d_i. The lower the perplexity value is, the better is the ability of the corresponding trained model in capturing the text collection.

Based on the proposed expanding approach, nodes in each layer are divided to subtopics where they contain sets of more compact health QA records as compared to their parents. As a byproduct of expansion, we train an optimal LDA model for each node in our generated hierarchy, which is utilized to facilitate health QA records assignment and hierarchy-based retrieval.

2.3 Shrinking approach

Before expanding a given node, we need to estimate how specific the node is, which is the key to automatically determining the depth of the hierarchy and prevents further segmentation of homogeneous nodes. Common approaches predefine a fixed depth and divide the data collection continuously until the depth constraint is satisfied. Approaches of this kind generate balanced trees where all leaves have the same depth. However, they have two limitations. First, the generated hierarchies might be biased towards the experiences of the persons who predefine the depth. Second, the underlying assumption of these approaches is that all sibling nodes have the same complexity and generality, which is not true in health domain. For example, the node talking about "cancer" is more general and should have deeper layers as compared to one that representing "acne."

We propose a shrinking approach to accomplish this task. Initially, we assume that the given node $\mathcal{V}$ can be further expanded, by dividing it into two child nodes, $\mathcal{A}$ and $\mathcal{B}$. Obviously, $\mathcal{V}$ equals to the union of $\mathcal{A}$ and $\mathcal{B}$, i.e.,

$\mathcal{V} = \mathcal{A} \cup \mathcal{B}$. We then estimate the average similarity between these two nodes by

$$R(\mathcal{A},\mathcal{B}) = \frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{\mathbf{x}_i \in \mathcal{A}, \mathbf{x}_j \in \mathcal{B}} S(\mathbf{x}_i, \mathbf{x}_j), \tag{2}$$

where $S(\mathbf{x}_i, \mathbf{x}_j)$ is their similarity estimation.

Based on the formulation of $R(\mathcal{A},\mathcal{B})$, we can intuitively have the normalized definitions of inter-node relation and intra-node relation as follows:

$$\begin{cases} \text{inter}(\mathcal{A},\mathcal{B}) = \frac{R(\mathcal{A},\mathcal{B})}{R(\mathcal{A},\mathcal{V})} + \frac{R(\mathcal{A},\mathcal{B})}{R(\mathcal{B},\mathcal{V})}, \\ \text{intra}(\mathcal{A},\mathcal{B}) = \frac{R(\mathcal{A},\mathcal{A})}{R(\mathcal{A},\mathcal{V})} + \frac{R(\mathcal{B},\mathcal{B})}{R(\mathcal{B},\mathcal{V})}. \end{cases} \tag{3}$$

The stronger the inter-node relation between $\mathcal{A}$ and $\mathcal{B}$ is, the more indivisible they are. On the other hand, a smaller intra-node relation indicates a more tighter consolidation of $\mathcal{V}$, and hence, it is not necessary to split it further.

In our work, if inter$(\mathcal{A},\mathcal{B})$ is larger than our threshold δ, we will terminate the expanding phase. The threshold is obtained empirically based on our experiments.

2.4 Health QA record assignment

As aforementioned, health QA records usually involve multiple topics. For example, this question is selected from HealthTap, "what can cause breast cancer to 25 years old married girl within the first 3 months of pregnancy?" It explicitly talks about at least three topics: "breast cancer," "female health," and "pregnancy." Therefore, assigning such records into multiple and complementary child nodes is desired in health domain.

Algorithm 1 The leading child nodes selection

```
1. Rank $p(\mathcal{V}_i|\mathbf{x})$ for different child nodes $\mathcal{V}_i$ in descending
   order.
2. Calculate the difference between two adjacent values in
   the ranking list.
3. Find the maximum difference, which is a boundary of
   leading child nodes and supporting child nodes.
4. Assign the given health QA records to the leading child
   nodes.
```

Based on our LDA model, each health QA record in the parent node can be represented as a mixture of all its children topics with different weights, i.e., $p(\mathcal{V}_i|\mathbf{x})$, denoting the probability of a health QA record $\mathbf{x}$ associated to a child node $\mathcal{V}_i$. Some child nodes with larger probabilities capture the principle components of the given health QA record, while others play supporting roles. However, there is not an indisputable approach to determine how many nodes should be selected for the assignment. If we choose too many child nodes, we may bring in noise for those nodes that are not the principle topics of the given health QA record. If we choose too few, we lose relevant category information of the given health QA record. As a rule of thumb, we should select only the leading interpretable child nodes.

An important observation reveals that the leading child nodes make significantly larger impact than other supporting child nodes. As Fig. 2 shows, there is a large gap between the impact of leading child nodes, i.e., v_1 and v_2, and that of the supporting child nodes, i.e., v_3, v_4, and v_5. This gap shows that the given QA record is highly relevant to the first two child nodes while it is less relevant to the last three child nodes. Hence, we assign the current health QA record, i.e., $\mathbf{x}$, into just highly relevant child nodes, i.e., v_1 and v_2 in our example. Based on this observation, the number of leading child nodes can be heuristically selected according to Algorithm 1. The algorithm first calculates the difference between two adjacent values in the ranked list of child nodes (line 2). It then finds the maximum difference to compute the number of leading nodes for current health QA records (line 3). The complexity of this algorithm is $O(nlogn)$. Similar approach was utilized to determine the leading roles from movies [11].

2.5 Node profiling with terminologies

Our LDA-based top-down scheme automatically extracts child nodes in the form of multinomial distributions of words from the parent node. In general, it is very difficult for users to understand a child node only based on the multinomial distribution of words, especially when they are not familiar with the context. Consequently, we need to generate meaningful labels for each node to ease understanding. In this section, we propose an approach for profiling nodes of the constructed hierarchy with medical terminology.

Early literatures [10, 12, 13] on topic labeling generally either select the top statistical terms in the distribution as primitive labels or generate labels manually in a subjective manner. These approaches, however, are not applicable to CHSs due to the following reasons. First, frequent terms might not be medical concepts, such as "desktop." Second, terms are less descriptive than phrase-based concepts. Third, manual generation is time-consuming and error-prone. In addition, terms are not standardized and inconsistent. Therefore, it is essential to automatically profile nodes with phrase-based standard terminologies.

Given one node, we initially assign part-of-speech tags to each word for all the health QA records associated with this node[8]. We then extract the noun phrases where their tag sequences match a fixed pattern,

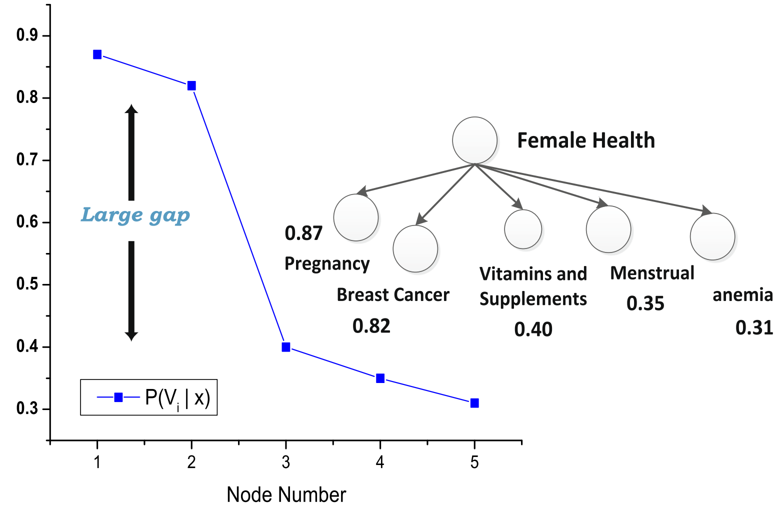

Fig. 2 Illustration of leading nodes selection

$$(\text{Adjective}|\text{Noun})^*(\text{Noun}\quad\text{Preposition})\;?(\text{Adjective}|\text{Noun})^*\text{Noun}. \tag{4}$$

A sequence matching this pattern ensures a noun phrase, such as the phrase "ineffective treatment of terminal lung cancer." We do basic post processing to link the variants of terms together, such as singularizing all plural variants.

We select the top k frequent noun phrases $\mathcal{C} = (c_1, c_2, \ldots, c_k)$ and normalize them into authenticated terminologies in ULMS via a voting method. More specifically, we first use MetaMap tool[9] to map each phrase into the ULMS terminology. It is worth highlighting that some distinct noun phrases may be mapped to the same terminology. For example, "painful neck" and "neck ache" are both normalized to "neck pain." We next use a voting strategy to rank terminology candidates $\mathcal{T} = (t_1, t_2, \ldots t_m)$ and produce the final labels by selecting the top ones,

$$\text{score}(t_i) = \sum_{j=1}^{k} \text{vote}(c_j, t_i), \tag{5}$$

where vote(c_j, t_i) is a binary form definition

$$\text{vote}(c_j, t_i) = \begin{cases} 1 & \text{if } t_i \text{ is terminology of } c_j \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

where Eq. (5) aggregates all the votes for each terminology phrase and Eq. (6) increases the score of a terminology if it can be inferred from a noun phrase.

A ranking list of terminologies for each node can be generated and the top ones are truncated as labels. The above voting approach preserves two characteristics. First, it assigns higher score to medical terminologies which are relevant to frequent occurring noun phrases in the cluster. Second, by inferring medical terminologies using MetaMap tool, we indeed normalize noun phrases into a standard medical terminology, i.e., UMLS.

3 Hierarchy-based retrieval

Reported by a national survey, which was conducted by the Pew Research Center[10], retrieval is the main mode of acquiring health information by users. Keyword-based indexing and matching is the prevailing method of retrieval. However, it is not sufficient for healthcare domain because of the complex, inconsistent and ambiguous terms used by users. In fact, the same questions may be described in substantially different ways by two individual health seekers, even by the well-trained doctors. For example, the query "I want to get pregnant what is the first thing I should do in diet and supplementary term?" and the archived health QA record "what are the best vitamins for a woman who decides to have a child soon?" are too semantically similar and both talking about mothers' worries about pregnancy. However, they are not very syntactically similar to be matched.

To boost the search performance, we propose a hierarchy-based retrieval application. It first deems the given query as a health QA record and performs health QA record assignment to the offline generated hierarchy. This is done by routing the given query from root level down to appropriate leaves of the tree. Obviously, this process plays an essential role in pruning the search space via routing the given query to the relevant branches. Meanwhile, the health QA record assignment actually employs the topic-based representation to semantically match the query to the relevant branches, which naturally tackles many of the limitations associated with term-based matching.

For a given query, a small set of leaf nodes are located. However, the health QA records within these selected leaf nodes are still large that will easily overwhelm the health seekers. Therefore, ranking these health QA records and returning the top ones to the health seekers will enrich the users' search experiences. The existing ranking approaches generally fall into two classes [14, 15]. One is pseudo relevance feedback based [16–18], which treats a significant fraction of the top documents as pseudo-positive examples and collects some bottom documents as pseudo-negative examples. They then either learn a classifier or cluster the documents to perform ranking. The other class is graph based [19–22] that propagates the initial ranking information over the whole graph until convergence. Inspired by [19], we adopt the graph-based random walk ranking method, which is formulated based on two assumptions:

1. The relevance probability function is continuous and smooth in semantic space. This means that the relevance probabilities of semantically similar health QA records should be close.
2. The final relevance probabilities should be close to the initialized ones for each health QA record.

We construct a graph where the vertices are health QA records and the edges reflect their pairwise similarities. We first introduce some notations. We use $\mathbf{W}$ to denote the initialed similarity matrix and W_{ij}, its (i,j)th element, indicates the similarity of $\mathbf{x}_i$ and $\mathbf{x}_j$, estimated using Eq. (13). Let d_{ii} denote the sum of the ith row of $\mathbf{W}$, i.e., $d_{ii} = \sum_j W_{ij}$. Then, the graph-based learning approach can be written as

$$\min_{\mathbf{y}} \frac{1}{2}\sum_{i,j} W_{ij}\left(\frac{y_i}{d_{ii}} - \frac{y_j}{d_{jj}}\right)^2 + \lambda \sum_i \frac{1}{d_{ii}}(y_i - \bar{y}_i)^2, \quad (7)$$

where λ is a weighting parameter and y_i is the relevance probability of $\mathbf{x}_i$ that we want to estimate. $\bar{y}_i$ is the initialized relevant score estimated by Eq. (13). We can see that the smoothness assumption is enforced in the first term of the above equation, which enforces the relevance probabilities of semantically similar health QA records to be close. The second term reflects the second assumption, i.e., the probabilities we estimate should be close to the ranking-based probabilities.

We use $\mathbf{D}$ to denote a diagonal matrix, with d_{ii} to be its (i,i)th element; and let $\mathbf{g}$ denote $\left[\frac{y_1}{d_{11}}, \frac{y_2}{d_{22}}, \ldots, \frac{y_n}{d_{nn}}\right]^T$. Thus, Eq. (7) can be rewritten as,

$$\min_{\mathbf{g}} \mathbf{g}^T(\mathbf{D} - \mathbf{W})\mathbf{g} + \lambda(\mathbf{g} - \mathbf{D}^{-1}\bar{\mathbf{y}})^T\mathbf{D}(\mathbf{g} - \mathbf{D}^{-1}\bar{\mathbf{y}}). \quad (8)$$

It can be derived that

$$\mathbf{y} = \frac{1}{1+\lambda}\mathbf{W}\mathbf{D}^{-1}\mathbf{y} + \frac{\lambda}{1+\lambda}\bar{\mathbf{y}}. \quad (9)$$

We can iterate the above equation and the convergence can be proven. With graph-based random walk ranking, we return an ordered list of health QA records to health seekers.

4 Features and similarity estimation

To represent QA records, we extract lexical, syntactic, and semantic features.

Weighted term kernel Φ_1: Medical concepts usually convey more informative signals than others. It is reasonable to assign greater weights to these concepts. We propose a weighted bag-of-word approach to lexically represent health QA content. Specifically, medical concepts falling into certain UMLS semantic groups will be weighted twice [23]. These groups include disease or syndrome, body part organ or organ component, sign or symptoms, and neoplasm. These groups are chosen since they cover most of the medical concepts and the medical concepts within them are discriminative. Cosine similarity is then employed to calculate the lexical similarity between two QA records.

Syntactic tree kernel Φ_2: The tree kernel function is one of the most effective ways to represent the syntactic structure of a sentence [24]. The tree kernel was designed based on the idea of counting the number of tree fragments that are common to both parsing trees, and defined as

$$STKN(T_i, T_j) = \sum_{n_i \in T_i} \sum_{n_j \in T_j} C(n_i, n_j), \quad (10)$$

where n_i and n_j are sets of nodes in two syntactic trees T_1 and T_2, and $C(n_i, n_j)$ equals to the number of matched sub-trees rooted in nodes n_i and n_j, respectively. STKN is originally designed to measure the similarity between two sentences. However, health QA records usually includes multiple sentences. We thus generalize it to Φ_2 as

$$\Phi_2 = \frac{\sum_{s_i \in d_1} \sum_{s_j \in d_2} STKN(T(s_i), T(s_j))}{|d_1||d_2|}, \quad (11)$$

where s_i and s_j are sentences from d_1 and d_2, respectively. In this way, we moderate the effects of the length of health QA records.

Latent topic kernel Φ_3: We explore the LDA-based high-level representation. For a collection of health QA records, LDA assigns semantically interrelated health concepts into the same latent group, which can be used to describe the underlying semantic structures of health data in the context of a hierarchical topic. In our work, each group is deemed as one feature dimension. Hence, for a given health QA record, it can be represented as a mixture of latent groups. The feature dimensions are determined via perplexity score.

Traditionally, the Kullback-Leibler divergence (KL-divergence) is used to compute the similarity between two topic distributions. However, KL-divergence is asymmetry per se, which makes it difficult to be used as a similarity metric. To address the asymmetry of KL-divergence, we utilize the Jensen-Shannon divergence scores as follows:

$$\Phi_3 = 0.5KL(p_1\|q) + 0.5KL(p_2\|q), \tag{12}$$

where $KL(.\|.)$ denotes KL-divergence score and $q = 0.5p_1 + 0.5p_2$ [25].

To estimate the similarity between two QA records, we linearly fuse these three aspects,

$$\Phi = \sum_{i=1}^{3} \beta_i \Phi_i, \tag{13}$$

where β_i sums up to 1, and each of them is greater than 0. We conduct a grid search with step size 0.05 within $[0, 1]$ to tune β_1 and β_2 while $\beta_3 = 1 - \beta_1 - \beta_2$. The values that achieved the best results are selected.

5 Experiments

5.1 Experimental settings

We collected approximately 109 thousand questions from HealthTap. For each question, we also collected its answers and tags, which are provided by doctors. Compared to normal documents, health questions are short and consist of only a few sentences. They thus do not provide sufficient word co-occurrences or shared contexts for effective similarity measurement. To compensate for this problem, we utilized corresponding answers and tags to contextualize the question parts. Note that for the hierarchy-based search, the newly incoming query contains only the question part.

For the subsequent subjective evaluations, we invited three volunteers who majored in medicine. They were trained with short tutorials and a set of typical examples before their labeling. A majority voting scheme among the volunteers was adopted to alleviate the problem of ambiguity.

5.2 On hierarchy generation

Currently, there are no widely accepted metrics to measure how well the generated hierarchy can explain the given data corpus. In our work, we propose objective and subjective approaches. We compare among three schemes: our scheme without domain knowledge, our scheme with domain knowledge, and hierarchical LDA (hLDA). The hLDA model [26] represents the distribution of topics within documents by organizing the topics into a tree. For hLDA, we assigned each health QA record into one child node based on the generative probability. We profiled each node with terminologies via mapping the top terms in each node to terminologies.

5.2.1 On objective hierarchy evaluation

We objectively evaluate the generated hierarchies from local and global angles. Both of these two evaluation approaches view external standard medical knowledge structure as golden hierarchies. In our work, we chose medical subject headings[11] (MeSH) as ground truth. It is a national library of medicines controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. At the most general level of the hierarchical structure are very broad headings such as "anatomy" or "mental disorders." More specific headings are found at the narrower levels of the 12-level hierarchy, such as "ankle" and "conduct disorder." There are 27,149 descriptors in *2014* MeSH.

For the local evaluation, we estimated the proportion of correct parent-child relations between labeled terminologies. We first formed a collection of relation tuples (parent, child) from the profiled hierarchies. We then inspected the correctness of each tuple in MeSH.

However, there exist some parent-child relations in our generated hierarchies which cannot be identified exactly in MeSH. For example, terminology t_i may be a grandchild of t_j in MeSH, while it is a child of t_j in the generated hierarchies. Therefore, local metrics are unable to comprehensively reflect the hierarchy cohesiveness. That motivates a global measure to estimate the cohesiveness,

$$\text{cohesivenss} = \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} R(t_i, t_j), \tag{14}$$

where t_i is the terminology in the parent node of the generated hierarchy, while t_j is the terminology in the adjacent child node. $R(t_i, t_j)$ is calculated based on MeSH,

$$R(t_i, t_j) = \begin{cases} \frac{1}{2^p} & \text{if ancestor-child relations} \\ 0 & \text{otherwise} \end{cases}, \tag{15}$$

where p is the length of ancestor-child path between terminology t_i and t_j.

The local and global evaluation results are presented in Table 1. It can be seen that our approaches extract much more relations between concepts from corpus. Moreover, our approaches outperform the hLDA in terms of local and global evaluations. The low cohesiveness values are caused by the fact that some parent-child terminologies are not represented in MeSH in the ancestral series. Finally, even though we use a very basic domain knowledge, it boosts the performance of the hierarchy generation, which validates the importance of domain knowledge for organizing medical data.

Table 1 Local and global evaluation results of the generated hierarchies

Approaches	Total tuples	Correct tuples	Accuracy (%)	Cohesiveness
hLDA	74	17	22.97	1.2×10^{-4}
Ours without domain knowledge	398	158	39.70	2.0×10^{-4}
Ours with domain knowledge	326	134	41.10	3.1×10^{-4}

5.2.2 On subjective hierarchy evaluation

As a complementary evaluation approach, we subjectively validated the generated hierarchies. We asked the volunteers to not only focus on the high-level parent-child relations in terms of labeled terminologies, but also the fine-grained context of the generated hierarchy. Because the hierarchies are very large, we first segmented each hierarchy into several tree fragments based on three conditions:

- Each fragment contains at most two levels.
- At most four siblings are allowed.
- Fifty records were randomly sampled from each selected node to represent its context.

The volunteers were required to go through all the health QA records in each fragment, which help them to grasp the contexts. After that, they were asked to annotate each fragment with ratings of "very satisfied," "satisfied," and "not satisfied." The results are presented in Table 2. As can be seen, our proposed schemes significantly outperform hLDA. Meanwhile, the hierarchy generated with domain knowledge can further reduce the "not satisfied" cases. We also evaluated the inter-volunteer agreement with the Kappa method [27]. The overall agreement value is 85.99%, while the fixed-marginal Kappa and free-marginal Kappa values are 0.7736 and 0.7899, respectively. They demonstrate that there are sufficient inter-volunteer agreements.

5.2.3 On health QA records assignment

Our scheme enables each health QA record to be assigned into multiple siblings. According to our statistics, on average each record is categorized into 1.7 child nodes. We aim to evaluate the precision and recall of our assignment approach. Precision equals to the number of correctly assigned child nodes over all assigned child nodes, while recall measures the fraction of principle topics of the given health QA record that are captured by the assigned child nodes. As aforementioned, for hLDA, each health QA record in the parent node was assigned into only one child node, which serves as a baseline to see how well our assignment approach performs.

Specifically, we randomly selected 20 nodes and their child nodes from each of the three hierarchies. For each node, we randomly sampled 10 health QA records. Three volunteers were first asked to go through each node and their child nodes to understand what subtopic each child node stands for. In fact, this stage provides cues to the volunteers to which child nodes the given health QA record should be assigned. Suppose the volunteer thinks that the given health QA record should be assigned into v child nodes, while it was only correctly assigned into u, then the recall for this health QA record is u/v. Average recall over three volunteers was calculated for each health QA record. Naturally, we also obtained the assigning precision for each health QA record. Table 3 presents the results. It can be seen that our schemes show superiority over hLDA. Our scheme with domain knowledge achieves promising performance in terms of recall.

5.2.4 On node profiling with terminologies

It is well known that for the labeling task, precision is usually more important than recall. We thus adopted two metrics that are able to characterize precision from different aspects. The first one is average $S@K$ over all testing nodes, which measures the probability of finding a relevant terminology among the top K recommended candidate terms. To be specific, for each testing node, $S@K$ is assigned to 1 if a relevant terminology is positioned in the top K terms and 0 otherwise. The second one is average $P@K$ that measures the proportion of recommended terminologies that are relevant. It is formulated as $P@K = \frac{|\mathcal{C} \cap \mathcal{R}|}{|\mathcal{C}|}$, where $\mathcal{C}$ is a set of top K terminologies and $\mathcal{R}$ is the manually labeled positive ones. The volunteers were required to label only top five suggested terminologies for each node, and they were labeled either as "positive" or "negative."

Table 2 Subjective evaluation of generated hierarchies

Approaches	# of fragments	Very satisfied	Satisfied	Not satisfied
hLDA	50	16	7	27
Ours without domain knowledge	50	34	10	6
Ours with domain knowledge	50	32	14	4

Table 3 Subjective evaluation of assignments of health QA records into hierarchies

Approaches	# of selected nodes	# of Sampled records	Recall (%)	Precision (%)
hLDA	20	200	48.2	76.5
Ours without domain knowledge	20	200	61.7	82.56
Ours with domain knowledge	20	200	65.86	82.33

Table 4 illustrates the results in terms of $S@K$ and $P@K$. It can be seen that our methods consistently outperform hLDA in both $S@K$ and $P@K$. This may be caused by the use of frequent terms in hLDA that are not medical terms.

5.3 On hierarchy-based retrieval

We comparatively evaluate the following unsupervised reranking methods:

- **KB**: term-based matching was implemented based on Apache Lucene[12] via indexing all health QA records in our data corpus.
- **PRF**: pseudo-relevance feedback [16].
- **R_noDK**: retrieval based on our scheme without domain knowledge.
- **R_DK**: retrieval based on our scheme with domain knowledge.

To obtain the relevance ground truth of returned health QA record, we conducted a manual labeling procedure. Each health QA record was labeled by three volunteers to be very relevant (score 2), relevant (score 1), or irrelevant (score 0) with respect to the given query. We adopted *NDCG@n* as our metric [28].

We randomly sampled 50 questions as queries. Figure 3 illustrates the experimental results with various *NDCG* depths. It can be observed that our proposed hierarchy-based retrieval approaches consistently outperform the other prevailing techniques. The possible reason may be the different search space. **KB** and **PRF** search over the entire data corpus, while ours route the given query to relevant leaf nodes that ensures the relevant search space in semantic topic level. The following graph-based random walk reranking further improves the precision. In addition, the **R_DK** approach performs better than **R_noDK**, because our scheme without domain knowledge is unable to precisely partition high-level groups.

6 Related work

Related literatures on organizing user-generated contents can roughly be classified into three categories: pattern-based, statistical, and folksonomy-based approaches.

The pattern-based approaches utilize predefined linguistic rules to identify concepts and their inter-relations, such as “is-a” and “whole-part.” For example, Li et al. [29] defined a subsumption relation to extract ontological relations between complex concepts from text segments. Beyond hierarchy generation on individual data source, the effort in [2] concentrated on organizing information resources into a topic hierarchy from multiple independent sources.

Statistical approaches either use hierarchical clustering methods or build a model to generate the hierarchy. For instance, Ming et al. [1] clustered web knowledge based on a predefined prototype hierarchy. Cimiano and Staab [30] constructed a hierarchy using agglomerative clustering and a hypernym oracle. Another example, Wang et al. [31] used generative model to cluster concepts for organizing information sources.

Folksonomy-based approaches attempt to generate hierarchies in lights of the collaborative annotated tags. Tang et al. [32] presented an ontology learning method using generative probabilistic model. Tsui et al. [33] used heuristic rules and a concept-relation acquisition schema to convert folksonomies to taxonomy. Song et al. [34] proposed a hierarchical tag visualization approach based on greedy algorithm. They then iteratively selected an optimal tag from the ranking list and inserted it into the tree following the minimum-evolution criteria.

However, most of these approaches are not suitable for CHSs due to the following issues. First, they usually allow each data instance to be assigned into only one leaf node. While each record in health domain usually covers more than one concern. Second, they label each node by a set of frequent concepts and terms instead of

Table 4 The evaluation results of node profiling with terminologies in terms of *S@K* and *p@K*

	S@1 (%)	S@3 (%)	S@5 (%)	P@1 (%)	P@3 (%)	P@5 (%)
hLDA	50	72	100	50	48.33	45
Ours without domain knowledge	52.5	80	100	52.5	50.83	46.5
Ours with domain knowledge	57.5	87.5	100	57.5	49.17	48

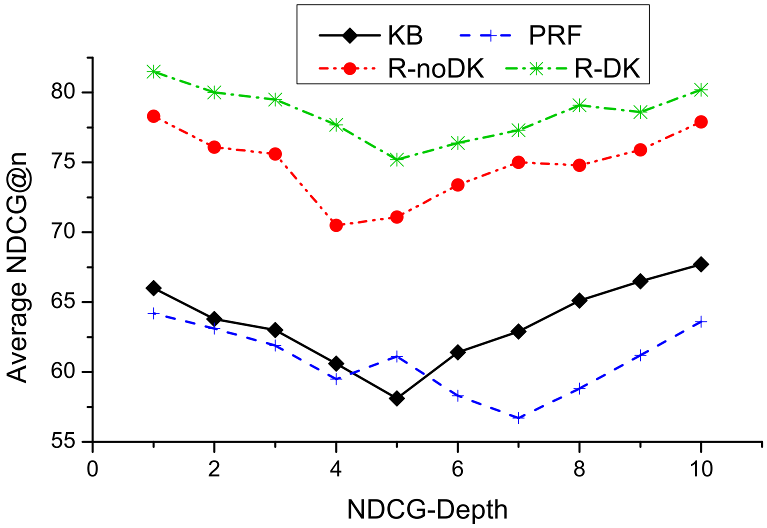

Fig. 3 Performance comparison among search algorithms in terms of NDCG@N

standard terminologies, which is not feasible for inter-system operations. Most importantly, the existing efforts do not consider flexible number of sub topics and layers for topic hierarchies.

7 Conclusions

This paper presented a novel top-down hierarchy generation scheme that is able to automatically organize the community-contributed health data with prior domain knowledge. Each node in the generated hierarchy was labeled with terminologies. Meanwhile, each health record can be categorized into more than one leaf nodes. Based on the generated hierarchy, a search function was designed and implemented to boost health information retrieval performance.

Our future work will focus on query-aware hierarchy generation. Specifically, given a natural language query, we will return a comprehensive hierarchy that covers various aspects expected by the query.

Endnotes

[1] http://health.yahoo.net

[2] http://www.webmd.com

[3] https://www.healthtap.com

[4] http://www.haodf.com

[5] http://sg.answers.yahoo.com

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

[7] http://www.healthxchange.com.sg

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

[9] http://metamap.nlm.nih.gov/

[10] http://pewinternet.org/Reports/2013/Health-online.aspx

[11] http://www.nlm.nih.gov/mesh/

[12] http://lucene.apache.org

Authors' contributions

All authors contributed equally in this work. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School for Integrative Sciences and Engineering, NUS, Singapore, Singapore. [2]School of Computing, NUS, Singapore, Singapore. [3]Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA.

Received: 16 June 2016 Accepted: 24 October 2016
Published online: 17 November 2016

References

1. Ming, Z-Y, Wang, K, Chua, T-S (2010). Prototype hierarchy based clustering for the categorization and navigation of web collections, In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 2–9): ACM.
2. Zhu, X, Ming, Z-Y, Zhu, X, Chua, T-S (2013). Topic hierarchy construction for the organization of multi-source user generated contents, In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 233–242): ACM.
3. Nie, L, Zhao, Y, Wang, X, Shen, J, Chua, T-S (2014). Learning to recommend descriptive tags for questions in social forums. *ACM Trans Inf Syst (TOIS)*, *32*(1), 5. ACM.
4. Golder, SA, & Huberman, BA (2006). Usage patterns of collaborative tagging systems. *J. Inform. Sci*, *32*(2), 198–208. Sage Publications.
5. Mankoff, J, Kuksenok, K, Kiesler, S, Rode, JA, Waldman, K (2011). Competing online viewpoints and models of chronic illness, In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 589–598): ACM.
6. Yang, S-H, White, RW, Horvitz, E (2013). Pursuing insights about healthcare utilization via geocoded search queries, In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 993–996): ACM.

7. White, RW, & Horvitz, E (2012). Studies of the onset and persistence of medical concerns in search logs, In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 265–274): ACM.
8. Cartright, M-A, White, RW, Horvitz, E (2011). Intentions and attention in exploratory health search, In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 65–74): ACM.
9. Luo, G, & Tang, C (2008). On iterative intelligent medical search, In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 3–10): ACM.
10. Blei, DM, Ng, AY, Jordan, MI (2003). Latent dirichlet allocation. *J Mach Learn Res*, *3*, 993–1022.
11. Weng, C-Y, Chu, W-T, Wu, J-L (2009). Rolenet: Movie analysis from the perspective of social networks. *IEEE Trans Multimed*, *11*(2), 256–271. IEEE.
12. Ramage, D, Heymann, P, Manning, CD, Garcia-Molina, H (2009). Clustering the tagged web, In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 54–63): ACM.
13. Mao, X-L, Ming, Z-Y, Zha, Z-J, Chua, T-S, Yan, H, Li, X (2012). Automatic labeling hierarchical topics, In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 2383–2386): ACM.
14. Nie, L, Yan, S, Wang, M, Hong, R, Chua, T-S (2012). Harvesting visual concepts for image search with complex queries, In *Proceedings of the 20th ACM international conference on Multimedia* (pp. 59–68): ACM.
15. Nie, L, Wang, M, Zha, Z, Li, G, Chua, T-S (2011). Multimedia answering: enriching text QA with media information, In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 695–704): ACM.
16. Liu, Y, Mei, T, Hua, X-S, Tang, J, Wu, X, Li, S (2008). Learning to video search rerank via pseudo preference feedback, In *2008 IEEE International Conference on Multimedia and Expo* (pp. 297–300): IEEE.
17. Natsev, AP, Naphade, MR, Teš,iĆ (2005). Learning the semantics of multimedia queries and concepts from a small number of examples, In *Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 598–607): ACM.
18. Yan, R, Hauptmann, A, Jin, R (2003). Multimedia search with pseudo-relevance feedback, In *International Conference on Image and Video Retrieval* (pp. 238–247): Springer.
19. Akbari, M, Nie, L, Chua, T-S (2015). aMM: Towards adaptive ranking of multi-modal documents. *Int J Multimedia Inf Retr*, *4*(4), 233–245. Springer.
20. Nie, L, Akbari, M, Li, T, Chua, T-S (2014). A joint local-global approach for medical terminology assignment, In *MedIR@ SIGIR* (pp. 24–27).
21. Nie, L, Zhao, Y-L, Akbari, M, Shen, J, Chua, T-S (2015). Bridging the vocabulary gap between health seekers and healthcare knowledge. *IEEE Trans Knowl Data Eng*, *27*(2), 396–409. IEEE.
22. Akbari, M, Huc, X, Liqianga, N, Chua, T-S (2016). From tweets to wellness: Wellness event detection from twitter streams, In *Thirtieth AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11931.
23. Sondhi, P, Sun, J, Zhai, C, Sorrentino, R, Kohn, MS, Ebadollahi, S, Li, Y (2010). Medical case-based retrieval by leveraging medical ontology and physician feedback: Uiuc-ibm at imageclef 2010, In *CLEF*.
24. Wang, K, Ming, Z, Chua, T-S (2009). A syntactic tree matching approach to finding similar questions in community-based qa services, In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 187–194): ACM.
25. Lin, J (1991). Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory*, *37*(1), 145–151. IEEE.
26. Blei, DM, Griffiths, TL, Jordan, MI, Tenenbaum, JB (2004). Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, *16*, 17. The MIT Press.
27. Warrens, MJ (2010). Inequalities between multi-rater kappas. *ADAC*, *4*(4), 271–286. Springer.
28. Nie, L, Wang, M, Zha, Z-J, Chua, T-S (2012). Oracle in image search: a content-based approach to performance prediction. *ACM Trans Graph (TOIS)*, *30*(2), 13. ACM.
29. Li, T, Chubak, P, Lakshmanan, LV, Pottinger, R (2012). Efficient extraction of ontologies from domain specific text corpora, In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1537–1541): ACM.
30. Cimiano, P, & Staab, S (2005). Learning concept hierarchies from text with a guided agglomerative clustering algorithm, In *ICML 2005 workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods, Bonn, Germany* (pp. 6–16): Citeseer.
31. Wang, C, Danilevsky, M, Desai, N, Zhang, Y, Nguyen, P, Taula, T, Han, J (2013). A phrase mining framework for recursive construction of a topical hierarchy, In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 437–445): ACM.
32. Tang, J, Leung, H-f, Luo, Q, Chen, D, Gong, J (2009). Towards ontology learning from folksonomies, In *IJCAI*, 9 (pp. 2089–2094).
33. Tsui, E, Wang, WM, Cheung, CF, Lau, AS (2010). A concept relationship acquisition and inference approach for hierarchical taxonomy construction from tags. *Inf Process Manag*, *46*(1), 44–57. Elsevier.
34. Song, Y, Qiu, B, Farooq, U (2011). Hierarchical tag visualization and application for tag recommendations, In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1331–1340): ACM.