On the Organization and Retrieval of Health QA Records for Community-based Health Services

Mohammad Akbari[†], **Xia Hu**[‡], **Liqiang Nie**[†], **Tat-Seng Chua**[†]
[†] National University of Singapore, [‡] Texas A&M University

1 Motivation

The urgent needs for online health information resulted in the establishment of many community-based health services such as HealthTap[1] and HaoDF[2]. They offer knowledge in the form of question answer (QA) pairs, where questions are issued by grassroot healthseekers, and answers are provided by doctors. These services have some intrinsic properties. First, they are crowdsourcing data that are continually growing at 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. Third, health seekers and doctors with diverse backgrounds present the same concepts in colloquial style, which leads to wide vocabulary gap. Together, these pose big challenges for data access and navigation. Recent efforts [Akbari *et al.*, 2016] indicate that organizing the community-contributed data into a hierarchical structure may enhance coarse-grained browsing and fined-grained search.

2 Method

In this work, we propose a top-down scheme that can organize the unstructured health QA records into a structured hierarchical tree. Our proposed scheme comprised of a pipeline with three main component: top-down hierarchy construction, QA record assignment, and hierarchy's nodes profiling.

In a topic hierarchy, nodes in higher layers of the tree represent abstract topics. These nodes usually hold a broad spectrum of subtopics, 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. To do so, we utilize the categorization of healthexchange[3] as our initial first layer following the root node. We then construct a set of classifiers to categorize health QA records in the root node into these categories, where we trained a SVM classifier for each category and obtained the training data through a pseudo-labeling approach extracting highly relevant QA records for each topic.

To construct the rest of hierarchy, we propose an expanding and shrinking approach to perform overlapping partitioning of each node to generate its children. Starting from the higher layer node, we assumed 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, where the proper number of children for each given node is computed via perplexity minimization. However, without termination criteria, the generated tree will be very huge in which each leaf nodes 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 expanding and shrinking approaches to generate a proper hierarchy.

To assign QA records to the nodes of the generated hierarchy, according to the LDA model, each health QA record in the parent node is 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 $\mathcal{V}_i$. Hence, child nodes with larger probabilities capture the principle components of the given health QA record, while others play supporting roles.

Finally all involved nodes are profiled with terminologies selected from Unified Medical Language System (UMLS) Metathesaurus, where a voting strategy is used to rank terminology candidates based on all QA records falling into the given node. Based on our proposed organization scheme, we develop a hierarchy-based QA retrieval system. Our application adopts the topic-based matching and performs intelligent pruning of irrelevant branches of the generated hierarchy.

3 Conclusion

We found that proper hierarchy construction, incorporation of domain knowledge, accurate record assignment, and node profiling can improve the final performance of the QA retrieval system with 7%-10% improvement over the baselines.

References

[Akbari *et al.*, 2016] Mohammad Akbari, Xia Huc, Nie Liqianga, and Tat-Seng Chua. On the organization and retrieval of health qa records for community-based health services. In *arXive*, 2016.

[1] https://www.healthtap.com
[2] http://www.haodf.com
[3] http://www.healthxchange.com.sg