Patient-Centric, Multi-Role, and Multi-Dimension Information Exploration on Online Healthcare Forums

Yunzhong Liu
Computer Science and Engineering
Arizona State University
Tempe, AZ, USA
liuyz@asu.edu

Yi Chen *
School of Management
New Jersey Institute of Technology
Newark, NJ, USA
yi.chen@njit.edu

ABSTRACT

Online healthcare forums provide a valuable platform for people to share medical information and support each other. However, currently the rich information shared on healthcare forums has not been fully explored. In this work, we first motivate the need for patient-centric, multi-role, and multi-dimension information exploration. We then present our patient-centric information exploration prototype system and show its effectiveness with preliminary experiment evaluation. We have also identified some potential techniques for multi-role and multi-dimension information exploration.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

Keywords

Information Exploration; Healthcare Informatics; Online Forums; Patient-Centric; Multi-Role; Multi-Dimension.

1. INTRODUCTION

Online healthcare forums are a very valuable platform for patients, caregivers, doctors, and researchers to share experience, seek insightful information, and support each other. Nowadays, there are many popular healthcare forums available, such as Patientslikeme [1] and WebMD [2]. They provide a large repository of user cases, evidences, and facts shared by patients or caregivers, which are very important resources for researchers to analyze various aspects of different diseases. Patients and caregivers can also seek help from other patients with similar symptoms or experts with abundant treatment experience.

However, currently the rich information shared on healthcare forums has not been fully explored. For example, the information search on most healthcare forums has low quality [9]. Assume a user wants to check if Vitamin can be used to alleviate aggression. She would like to issue a multi-keyword query "Vitamin, aggression" on a healthcare forum. Post-based search has been commonly used in many forums, such as the Patientslikeme and WebMD forum. It returns a post if it contains all the queried keywords. The post-based search may miss some relevant results and thus suffers low recall when the queried keywords come from multiple posts. For example, one caregiver may initiate a thread and describe some symptoms or behaviors of a patient she cares about. While a follower may comment in a replying post that those are a type of "aggression", another experienced forum user may further reply and suggest the patient to take "Vitamin". All this information centered around the patient answers the query perfectly but is missed by the posted-based search. To improve the recall, we can adopt the thread-based search, which and its variants are also used in many forums, such as the Healthboards message boards [3] and the Epilepsy forum [4]. It returns a thread or its link if all the posts in this thread collectively contain all the queried information. But the thread-based search suffers low-precision as shown next. Assume a user suffers from seizures due to weaning, and would like to know how other similar patients deal with this problem. She could search the forum by issuing a query "seizure, wean". Suppose there is a thread with multiple caregivers or patients participating in the discussion. One caregiver mentions her mother has seizures while another caregiver mentions her son is weaning off a medicine. This thread will be returned, but it is not very relevant to the user's search intention since it does not include the information about the seizures caused by weaning. The user may expect the returned disease, symptom, and their cause or treatment information are experienced by or associated with the same patient. Therefore, to effectively support such applications, a patient-centric information organization is critical, in which all the shared information related to the same patient should be identified and aggregated together for patient-centric information exploration.

Multi-role and multi-dimension information exploration can further facilitate some important applications on healthcare forums. On healthcare forums, people with multiple roles, such as patient, caregiver, or expert, participate in the discussion or are mentioned in the posts. Note that the same person can be a patient for one disease or issue, but can also be an expert for another one. If we can identify these persons and their roles in different contexts, it would be useful to fully leverage their expertise for helping each other.

On the other hand, multi-dimension information exploration categorizes the shared information on the forum from different per-

*Yi Chen is the PhD advisor of Yunzhong Liu.
[1]http://www.patientslikeme.com
[2]http://exchanges.webmd.com

PIKM'13, November 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2513166.2513174 ...$15.00.

[3]http://www.healthboards.com/boards
[4]http://epilepsyfoundation.ning.com/forum

spectives. For example, we can first categorize the medical information according to different diseases or issues, so that a forum user can easily find the other forum participants who share the same interests. Although a high-level disease categorization may have been provided by the forum itself, there are still various types and subtypes of the same disease. For example, epilepsy has two major types: generalized epilepsy and partial epilepsy. Each type also has many subtypes. They can have different symptoms and treatment methods. Such a fine-grained categorization can help the users to retrieve the most relevant information. For a type or subtype of disease, we can further categorize the information into different aspects, such as condition, symptom, cause, treatment, side effect, and etc. With such a categorization, we can have more flexible and desirable information search. For example, we can search patients with perfect symptom matching but with diverse treatment or side effect information. From another perspective, we can also categorize the information according to the information type, such as description information or suggestions, facts or opinions, and etc. With such a categorization, a query user can specify which type of information is needed.

With the patient-centric, multi-role, and multi-dimension information exploration, we may provide useful information statistics to researchers or forum coordinators. Given a sufficiently large amount of forum data available, we can collect some statistical information about the relationships between multiple diseases, the relationships between diseases and symptoms, the relationships between treatment and side effects, and etc. Such statistic information, summarized from individual patient cases, can bring new insights to medical research. We can also provide the distributions of description information and suggestions, the distributions of patients, caregivers, or experts for each different disease, and etc. Such information would be useful for the forum coordinators. For example, they may need more experienced experts for some specific diseases or issues.

With the above motivation, we have developed a patient-centric information exploration prototype system and proposed some approaches to improve the performance of system components. The preliminary experiment evaluation verifies the effectiveness of our system and approaches. We also propose to investigate how to leverage two probabilistic graphic models for further patient-centric, multi-role, and multi-dimension information exploration.

Related Work: There is some related work on medical information extraction, categorization, or integration on forums or other social media, such as [4, 11, 12, 14]. In [12], statistic machine learning techniques are used to classify the content of each sentence in forum posts into three types: symptom, treatment, or others. In [11], multi-dimension topic model is used to categorize the forum information according to the drug type, route of intake, and aspect. Their information extraction or categorization process is not patient-centric. In [4], the health data from different online communities are linked to provide integrated knowledge of health information. In [14], the authors present an architecture for personalized health information retrieval, in which the profile data of a query user (a patient) is utilized to selectively retrieve the relevant medical information. Different from our work, their information integration or retrieval relies on an existing user or patient profile.

2. PATIENT-CENTRIC INFORMATION EXPLORATION

In this section, we briefly introduce our patient-centric information exploration system with a preliminary experiment evaluation.

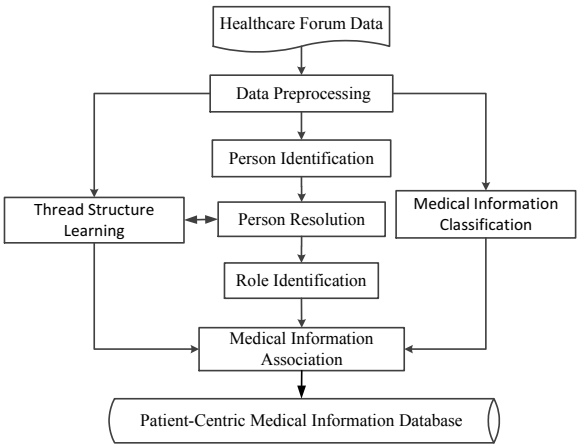

Figure 1: The system overview.

2.1 System Overview

Our patient-centric information exploration system includes seven major function components. Fig. 1 shows the whole process with the healthcare forum data as the input and the patient-centric medical information database as the output. With the original forum data, we first extract the information we need in the *data preprocessing* component, such as post title, author, content, posting time, post reply relationships within a thread, and etc.

To build a patient-centric medical information database, we need to identify all the patient mentions that refer to the same patient, and then associate and aggregate the medical information with the corresponding patients. Since a patient can be a post author or a person mentioned in a post, in the *person identification* component, we identify all the person mentions to find the potential patient mentions. While it is easy to identify the role of some person mentions from the sentences where they appear, it is difficult to identify the role of the others. For example, from the sentence "My daughter was diagnosed last Friday", we can identify "My daughter" is a patient by using semantic role labeling (SRL) [6] and a few heuristic role identification rules. But from the sentence "She is 5 years old", we cannot identify the role of "She". Therefore, after identifying all the person mentions, in the *person resolution* component, we first group all the person mentions into clusters such that all the mentions in the same cluster refer to the same person. In the *role identification* component, we then identify the role of each person based on the information in each cluster. The thread structure, the reply relationships between posts, is also important in our system. For example, the reply structure can be used for person resolution and they can boost the performance for each other, which will be explained in Section 3. However, such a reply structure may not be known or is only partially known if some forum users do not specify the reply relationship explicitly. The *thread structure learning* component is used to learn a complete thread reply structure. We also have the *medical information classification* component for categorizing the forum information. In our current system, we use MetaMap tool [2] to extract medical phrases and classify them into different semantic types. At last, in the *medical information association* component, we associate the medical information with the corresponding patients. Currently, we associate the medical information with the closest patient. The thread reply structure is also taken into consideration when making association.

2.2 Preliminary Experiment Evaluation

We evaluate our system with the search application on the epilepsy foundation forum data [9]. In this experiment, we assume that a

Table 1: Performance comparison on multi-keyword queries.

Post-based			Thread-based			Patient-centric		
P	R	F1	P	R	F1	P	R	F1
0.995	0.326	0.491	0.417	**1.0**	0.589	0.851	0.674	**0.752**

post with an unknown parent replies to the first post and the thread structure has been used for person resolution. We randomly chose ten thread titles that are a question, and extract medical phrases as keywords to form the queries. For example, we extract "Vitamin, aggression" as a multi-keyword query from the question title "Vitamins to help with aggression?" We then compare the query quality based on the original forum data with that based on our patient-centric database. As we describe in Section 1, we assume a user expects all the matched keywords in each returned result are related to the same patient. We assume a post-based search returns all the posts each containing all the query keywords. A thread-based search returns all the posts each containing at least one query keyword in a relevant thread, which contains all the query keywords. Our patient-centric search returns all the posts each containing at least one query keyword associated with a relevant patient, whose associated medical information matches all the query keywords.

Table 1 shows the average precision (P), recall (R), and F1-measure. Precision is the ratio of the number of correctly returned posts to the total number of returned posts. The recall is the ratio of the number of correctly returned posts to the total number of posts that should be returned according to the user expectation. F1-measure is defined as the harmonic mean of precision and recall: $F1 = \frac{2*P*R}{P+R}$. The experimental results show that post-based approach has almost perfect precision as in most cases keywords in the same post refer to the same patient, but it has very low recall. On the other hand, thread-based search achieves perfect recall since we do not consider the relationships of keywords in different threads in this experiment, but it has a very low precision. In contrast, our patient-centric search has good precision and recall in general, and achieves a much higher F1-measure than the other two approaches.

3. PERSON RESOLUTION AND THREAD STRUCTURE LEARNING

Person resolution (PR) is very critical during patient-centric information exploration, while the unknown thread reply structures also affect our system performance. For example, if the mentions of two different patients are wrongly identified as referring to the same person, the medical information associated with this person will be a mixture from two real patients, which will decrease the precision of our patient-centric search application in Section 2.2. On the other hand, if two mentions referring to the same patient cannot be identified and clustered, the patient's medical information will be divided and associated with two different persons, which will decrease the recall in the same application. The unknown thread reply structures not only make it difficult to do person resolution, but also directly affect the performance of medical information association. For example, a suggestion post without specifying the receiver is usually for the author of the parent post or the patient mentioned in the parent post. Without knowing which post is its parent, we cannot associate the suggestion information with its receiver. In this section, we focus on person resolution and thread structure learning and show how they can improve each other.

3.1 Thread Structure for PR

We first observe that the thread reply structure can be leveraged for inter-post person resolution within a thread. In general, one person mention can only refer to the person introduced before it in the same post or in its parent or ancestor post. While we can use the Stanford deterministic co-reference resolution system [1] for generating the co-reference resolution results in a post, we have to design our own inter-post person resolution system.

With the thread reply structure, we can do person resolution along the path from the thread root to the current post. In particular, we use the following matching rules to identify a pair of person mentions that refer to the same person within a thread: (1) Matching between the address in the current post content and the signature in the content of its parent post. Note that an address or a signature is usually a person's name or nickname, or the user ID of a post author. (2) Matching between the same role related to the same person. For example, "your daughter" in the current post can match "my/our daughter" in the parent post. (3) Matching between the second person pronoun in the current post and the first person pronoun in the parent post. (4) Matching between the third person pronoun in the current post and the person name or role in its parent post that are consistent in gender. These matchings are extracted by some NLP tools, such as [1], MetaMap [2], and some regular expressions.

3.2 PR for Thread Structure Learning

Most of the online healthcare forums, such as WebMD, Healthboards message boards, and the epilepsy forum used in our experiments, only have partially labeled thread reply structures. On these forums, there are always some post authors who have a good habit of keeping an explicit reply structure, while others do not. Our goal is to learn a complete thread reply structure with the partially labeled data.

While a clear thread reply structure can help identify the correct person reference relationships, a clear person reference relationship can also help learn the correct thread structure. For example, if one post mentions a person that is described in a preceding post, then this post is likely to be a child or descendant of that preceding post. According to this observation, if we can find out the person references in a person resolution process, that can provide helpful information for learning the thread reply structures.

We propose to use PR for thread structure learning [8]. Similar to the matching rules defined in Section 3.1, we define four types of PR features and each type has a different priority. Given a post, for each candidate parent post, we extract all the PR features, and then select out the candidate parent matched with the features of higher priority.

We can also combine PR with some statistic machine learning techniques to leverage the partially labeled data for inferring the unknown thread reply structures. Thread conditional random field (threadCRF), proposed in [13], has been shown to be very effective for thread structure learning. However, note that threadCRF is a supervised learning model. Before it can be applied for effective thread structure prediction, it requires a completely labeled data set for model training. In our combined approach, we use PR to generate a fully labeled training set given the partially labeled data for training threadCRF, and then use the learned threadCRF model to re-label those unknown reply relationships.

We refer our two methods described above as PR and PR + threadCRF, and compare them with three baseline methods: reply to the first post, reply to the last post, and reply to the post with the highest content similarity. These three baseline methods are referred as FIRST, LAST, and SIM, respectively. Table 2 shows the performance comparison among the five methods for learning thread structures of 200 selected threads on the epilepsy forum. The accuracy is defined as the proportion of correct labels in the whole set of predicted labels (468 unknown reply relationships in

Table 2: Thread structure learning performance.

	FIRST	LAST	SIM	PR	PR + ThreadCRF
Accuracy	0.444	0.429	0.361	0.583	**0.635**

this experiment). Among the first four methods, PR achieves the best performance. When we combine PR with threadCRF, we can see that the performance is significantly improved.

4. MULTI-ROLE AND MULTI-DIMENSION INFORMATION EXPLORATION

Multi-role and multi-dimension information exploration is related to our role identification and medical information classification components in our system. In addition to the integration of state of art NLP techniques and MetaMap tool, our previous work has mainly been focused on person resolution and thread structure learning, which are related to structure or link analysis. However, role identification and medical information classification, which are related to content or topic analysis, are also very important in our system. Furthermore, structure and content analysis can reinforce each other. For example, multi-role identification can be used to improve person resolution. In particular, if we assume one person's role does not change frequently in a local context, the role consistency can be used to identify if two person mentions refer to the same person.

Although some existing NLP techniques, such as semantic role labeling (SRL) [6], and MetaMap tool [2] have been used for role identification and medical information classification in our previous work, they cannot fully satisfy our application requirements. Previously, we mainly rely on some patterns or rules identified in a sentence combined with SRL to identify the role of a person, which is a deterministic approach and also fails to take the topic-level information associated with each person as the context into consideration. As we introduce in Section 1, the same person's role may change in different contexts. We may need a probabilistic approach to capture the role distribution of the same person in various contexts. While MetaMap tool can extract phrases from the text and classify them into different semantic types, such an information classification is along one dimension from the medical perspective, which cannot support the multi-dimension information exploration applications described in Section 1.

We will investigate the suitability of using probabilistic graphic models, such as conditional random field (CRF) [7] and topic modeling [3], for our multi-role and multi-dimension information exploration. Conditional random field has been used in our system for thread structure learning. It can also been used for role labeling [5] and medical information classification [12]. To apply CRF to our problem, the key challenge is to design the domain-specific feature set. On the other hand, topic modeling has been extensively used for text analysis. There are many topic model variants, including the recently proposed multi-dimension topic model [10]. Based on the work in [10], we can model the person-role-topic distribution, where a person will be associated with a sparse distribution over roles and a person-role pair will be associated with a multi-dimension topic distribution. Different from the existing work, both probabilistic models will be specifically designed for healthcare forums, and built on the other components in our patient-centric system.

5. CONCLUSIONS

In this paper, we propose patient-centric, multi-role, and multi-dimension information exploration on healthcare forums. We have developed a patient-centric information exploration prototype system and proposed some approaches to improve the performance of two important components. Some preliminary experiment evaluation verifies the effectiveness of our system and approaches. We also propose to investigate the suitability of using probabilistic graphic models for further patient-centric, multi-role, and multi-dimension information exploration.

Acknowledgments

This material is based on work partially supported by NSF CAREER Award IIS-0845647, IIS-0915438, an IBM Faculty Award and a Google Research Award. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

6. REFERENCES

[1] Stanford core NLP tools. `http://nlp.stanford.edu/software/corenlp.shtml`.

[2] A. R. Aronson. Metamap: Mapping text to the UMLS metathesaurus (2006). `http://skr.nlm.nih.gov/papers/references/metamap06.pdf`.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] S. A. Chun and B. MacKellar. Social health data integration using semantic web. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC)*, 2012.

[5] T. Cohn and P. Blunsom. Semantic role labeling with tree conditional random fields. In *Proceedings of CoNLL*, 2005.

[6] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

[7] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of International Conference on Machine Learning (ICML)*, 2001.

[8] Y. Liu, F. Chen, and Y. Chen. Learning thread reply structure on patient forums. In *Proceedings of International Workshop on Data management & Analytics for Healthcare*, 2013.

[9] Y. Liu and Y. Chen. Patient-centered information extraction for effective search on healthcare forum. In *Proceedings of International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction (SBP)*, 2013.

[10] M. J. Paul and M. Dredze. Factorial LDA: Sparse multi-dimensional models of text. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[11] M. J. Paul and M. Dredze. Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. In *Proceedings of NAACL-HLT*, 2013.

[12] P. Sondhi, M. Gupta, C. Zhai, and J. Hockenmaier. Shallow information extraction from medical forum data. In *Proceedings of International Conference on Computational Linguistics (COLING)*, 2010.

[13] H. Wang, C. Wang, C. Zhai, and J. Han. Learning online discussion structures by conditional random fields. In *Proceedings of ACM SIGIR*, 2011.

[14] N. Yadav and C. Poellabauer. An architecture for personalized health information retrieval. In *Proceedings of International Workshop on Smart Health and Wellbeing*, 2012.