Patient-Centered Information Extraction for Effective Search on Healthcare Forum

Yunzhong Liu and Yi Chen

School of Computing, Informatics, and Decision Systems Engineering
Arizona State University, Tempe, AZ
{liuyz,yi}@asu.edu

Abstract. Online healthcare forums are one of the major social media in Health 2.0 for patients and caregivers to share personal experience and to help each other. However, current forums do not support effective information search and thus users are unable to fully leverage the rich information in the forums. In this work, we propose patient-centered information extraction to better organize the information in the forum and have developed a patient-centered medical information database extracted from a forum. In this system, the patients discussed on the forum are identified and their shared medical information is aggregated and associated with the corresponding patients. The experimental evaluation shows that our system can provide better information search results than traditional approaches.

1 Introduction

Nowadays, Health 2.0, the web-based applications and services for healthcare, has become very popular. In Health 2.0, forums are one of the major social media where patients or their caregivers share personal experience, support and encourage each other, and form patient communities. In a forum, a user, or a post *author*, may publish a *post*, the smallest information unit in a forum. An initial post and the replying posts submitted by the same or different authors compose a *thread*, or a topic.

Online healthcare forums provide valuable information for patients, caregivers, doctors and researchers. There is a large and increasing volume of user cases, evidences, and facts shared by patients, which may provide insights to the research on diseases and treatments. It is also an important resource for patients and caregivers to seek for other patients with similar symptoms and to check what treatments have been taken by or suggested for those patients for self-education on their diseases and treatments.

However, currently the rich information on healthcare forums has not been fully leveraged. While it is easy to share information by posts, and to browse and read the posts shared by other patients, current technology does not provide effective ways for a user to easily *discover* information that she is interested in, in a large repository of posts. Let us look at two examples, both of which are observed in questions issued by real users to the epilepsy discussion forum[1].

[1] http://epilepsyfoundation.ning.com/forum

A.M. Greenberg, W.G. Kennedy, and N.D. Bos (Eds.): SBP 2013, LNCS 7812, pp. 175–183, 2013.

Consider a user who wants to check whether Vitamin can be used to alleviate aggression, and would like to search the epilepsy forum for other patients' experience to gain more knowledge. She would issue a keyword query "Vitamin, aggression" on the forum. One approach commonly used in forums to support information search is to consider each post as an information unit (like a document) and to return a post if it contains the query keywords, referred as *post-based search* in this paper. The Patientslikeme forum[2] and WebMD forum[3] are mainly based on this method.

Adopting the post-based search, the information shown in Table 1 will be missed from the result since there is no single post in this thread containing both query keywords. For space reason, only post fragments are included in the table. PostID is a post's sequence number in a thread. For privacy concern, we replaced the real AuthorID of the forum participants with C1, C2, and C3. The ParentPostID is the PostID of the post that the current post replies to. For example, the 4*th* post with PostID 4 replies to the first post with PostID 1. However, when we read through posts 4 and 6 in this thread, we can see that even though they are from different authors, they are closely connected and collectively show that Vitamin B6 can help aggression, which could be caused by Keppra.

As we can see, post-based search tends to put too strict criteria on search and thus misses relevant results, that is, it suffers *low recall*. To improve recall, an intuitive approach is to search forums using *thread-based search*: take each thread as an information unit and consider a thread as relevant if all the posts in the thread collectively contain the query keywords. This approach and its variants have been used in the Healthboards message boards[4] and the Epilepsy forum. Thread-based search can identify the thread in Table 1 as relevant to query "Vitamin, aggression". However, it suffers other problems, as shown in the following example.

Suppose a patient suffers seizure due to weaning, and would like to search the epilepsy forum to learn how to cope with her problems from similar patients. She would issue a keyword query "seizure, wean". Using the thread-based search method, a thread with title "B6 wondering" will be returned, where some fragments are shown in Table 2.

As we can see, this thread is returned as a query result since it contains a post with PostID 6 and a post with PostID 11, which together contain both query keywords. However, after reading these two posts, we find that keyword "seizure" and "wean" are associated with different patients: C4's mom and C5's son. There is no relationship between "seizure" and "wean" described in the thread, and thus this thread is not useful for the user who searches for the information about the seizure disease caused by weaning. As illustrated in this example, the thread-based search tends to return some results that are not relevant to the user, that is, it suffers *low precision*.

[2] `http://www.patientslikeme.com`
[3] `http://exchanges.webmd.com`
[4] `http://www.healthboards.com/boards`

From the above examples, we observe that existing approaches do not perform well for a user query with multiple keywords. We analyze those queries and find that when a query contains multiple keywords, these keywords are expected to have close relationships between each other. For instance, a query may involve the relationship between a symptom and a disease, the relationship among several symptoms, the relationship among multiple diseases, the relationship between a disease and treatments, or the relationship between a treatment and side effects. To correctly find such relationships, it is critical that the matches to query keywords refer to the *same* patient. However, post-based or thread-based search does not consider *who* a keyword is associated with. They only check syntactic information units, either a post or a thread. It is common to see multiple posts refer to the same patient, and a thread contains information of multiple patients. Therefore the root cause of the low-quality results generated by existing approaches is the mis-alignment between the syntactic information unit (a post or a thread) that existing methods are based on and the semantic information unit (a patient) that the query user refers to.

Table 1. Samples from one thread for the query "Vitamin, aggression"

Thread link: http://epilepsyfoundation.ning.com/forum/topics/katies-temper			
PostID	AuthorID	*Content*	Parent PostID
1	C1	Katie woke up a swinging her arms this morning and hitting things,very combative. I just wished I knew if it was the Keppra she is taking that is making her do this. She is VERY tempermental,alot of times I don't know what to do with her. ... She has in home therapys and her therapists the other day was telling me it seems she has a sensory integration disorder.	Null
4	C2	Yes, Keppra can cause **aggression**.	1
6	C3	Have you tried giving her **Vitamin** B6 with the keppra?? It is supposed to help with the Keppra-rage.	1

Table 2. Samples from one thread for the query "seizure, wean"

Thread link: http://epilepsyfoundation.ning.com/forum/topics/b6-wondering			
PostID	AuthorID	*Content*	Parent PostID
6	C4	My mom is 59 and she takes keppra as well says that she gets tired very very early, usually around 7 she's just about ready for bed. She swears by keppra for controlling both her **Seizure**s and the auras.	1
11	C5	My son is **wean**ing off keppra, but he's still taking 250mgtwo times a day. (He was on something like 1000mg and life was hell). He gets angry really fast- right after taking his meds.	1

Table 3. Simplified records from the patient-centered medical information database

PatientID	Note	Medical Info ID	Medical Information
1	Katie	1	Katie woke up...
1	Katie	2	Yes, Keppra can cause ...
1	Katie	3	Have you tried giving her ...
2	C4's mother	1	My mom is ...
3	C5's son	1	My son is ...

In light of this observation, we propose to mine the semantic information unit - each individual patient and the associated information - from the posts. Then a user query is processed with respect to the semantic information unit, finding out the patients whose experience is related to query keywords and therefore can bring insights about the relationships among the keywords. We developed an information extraction system, which takes the original forum data as the input, identifies the patients and the information associated with each individual, and outputs a patient-centered medical information database. Table 3 shows a simplified version of our database records extracted from the information shown in Table 1 and Table 2. In Table 3, multiple pieces of information from three different posts in Table 1 are identified to be associated with the same patient. On the other hand, the information from post 6 and 11 in Table 2 is extracted and associated with two different patients with different PatientIDs. With such a patient-centered database, it is easy to find which patients are relevant to a user query, thus improving the search quality achieved by post-based or thread-based search. In our example, patient 1 corresponds to a relevant result to user query "Vitamin, aggression", while none of them is relevant to user query "seizure, wean".

Related Work: There are existing studies [8,4] on improving the traditional thread or post-based search on other types of forums, such as technical forum. However, these types of forums are different from healthcare forum since their focus is topics rather than individuals described in the posts. Although the thread structure information, such as the reply relationship, has been exploited in these studies, they do not make deep NLP analysis to mine the semantic information unit in the posts, like each individual patient in a health forum. Therefore, they would have similar problems as post-based or thread-based search for the two example queries discussed earlier.

2 System Overview

To build the patient-centered medical information database, we need to identify the patient mentions that refer to the same person and to associate and aggregate the medical information with the corresponding patients. Our system includes four major components. In *person identification* module, we discover all the person mentions to find the potential patient mentions. Since it is difficult to identify a patient from some individual person mentions, we apply *person resolution* to group all the person mentions into clusters such that all the mentions in the same cluster refer to the same person. Then we make *patient identification* based on all the information in each cluster of person mentions. At last, we make *medical information association* for each identified patient from the posts. State-of-art natural language processing (NLP) techniques and MetaMap tool [2] in the Unified Medical Language System (UMLS) [6] have been integrated into our system.

2.1 Person Identification

This component takes sentences in posts as input and outputs person mentions. Our method is based on the Stanford NLP [1] and MetaMap tool in UMLS. First, all the person names identified by Named Entity Recognition (NER) and pronouns (except "it") identified by Part of Speech (POS) tagger are identified as person mentions. For example, "Katie" and all "she" and "her" in post 1 in Table 1 are identified as person mentions. Second, all the phrases extracted by MetaMap with their semantic types belonging to "living beings" semantic group will be identified as person mentions. For example, "son" in post 11 in Table 2 can be identified as a person mention since its semantic type is "family group", which belongs to "living beings" semantic group.

2.2 Person Resolution

This component groups the person mentions within a thread into clusters such that each cluster includes all the mentions that refer to the same person. Stanford deterministic co-reference resolution system [5], which was the top ranked system at the CoNLL-2011 shared task, is used for generating the co-reference resolution results. Since co-reference resolution is more general than person resolution, we can easily extract the person resolution results from the co-reference resolution results. For post 1 in Table 1, "Katie" and all "she" and "her" in this post will be identified as co-referent.

In addition to person resolution within a post, we also incorporate the author information and the reply relationship between posts for inter-post person resolution. First, we assume the same role with the same relationship with the same author in the same thread refers to the same person. For example, if "my son" has been mentioned by the same author in two different posts in the same thread, we consider them as co-referent. Second, we transform one thread into multiple multi-person conversation documents based on the reply relationship, in which a post author is a speaker and the post content is analogous to the utterance. In this way, the person mention in the replying post that refers to the person in its parent post can be identified.

2.3 Patient Identification

This component identifies the patient mentions from the identified person mentions. We assume a person mentioned in a thread is either a patient or a non-patient. Then we propose to combine the semantic role labeling (SRL) [3], MetaMap, and a few patient identification patterns. We identify patients mainly using SRL with Propbank [7] annotation. In addition, we also used 12 patient identification patterns based on a sample data set. As shown in experimental evaluation later, this small number of patterns, such as "take *pharmacologic substance*", "have *disease* or *syndrome*", have a very high coverage in identifying patients and scale well in a large dataset. Here "*pharmacologic substance*" and "*disease* or *syndrome*" are two semantic types for medical phrase, which

can be extracted from post content by MetaMap. For example, in post 1 in Table 1, from "she has a sensory integration disorder" we can identify "she" is a patient since "sensory integration disorder" has the semantic type "*disease* or *syndrome*". Note that all the co-referent person mentions will be identified as patient mentions if at least one of them has been identified as a patient mention. Therefore, "Katie" and all "she" and "her" in this post will be identified as patient mentions.

2.4 Medical Information Association

This component associates the medical information with the closest patient or person if no patient has been identified at all. Note that the medical information in a replying post can also be associated with the patient mentioned in its parent post if that replying post does not introduce a new patient that is closer to the information. In Table 1, "aggression" in post 4 and "Vitamin B6" in post 6 are both associated with "Katie" in post 1. Also note that no information should be associated with "you" in post 6 or "I" in post 1 since they refer to the caregiver "C1", rather than a patient.

3 Experiments

To evaluate our system, we use the publicly available data in the epilepsy foundation discussion forum, which is initiated and maintained by National Institute of Neurological Disorders and Stroke (NINDS). We collected 9210 posts included in 911 threads (topics) published on the "Patient help patient" sub-forum by Nov. 2011. In this forum, the explicit quotation information has been used to identify the reply relationship between two posts. Otherwise, by default, we consider all the following posts in a thread reply to the first post in this thread, which follows the assumption in the feature used in [9] that the following posts tend to reply to the first one.

3.1 Query Set

Our query set includes ten multiple-keyword queries. In order to leverage real user queries without introducing bias, we follow the method used in [4] to randomly select queries. First, we find all the thread titles in the forum that end with a question mark. Since such a title indicates that a user, the thread initiator, is looking for answers to a question, it naturally represents as a user query. We then extract keywords from these thread titles. Instead of using a stopword list to filter out unimportant words, we choose MetaMap tool to extract phrases as the query keywords. The reason is that we want to identify each medical phrase containing multiple words and treat it as a unit in query processing. We randomly chose ten such thread titles with each corresponding to one query. We only tested ten queries because it is extremely labor-intensive to generate the ground truth for each query, especially since some queries may involve

a large number of threads, which may include an enormous number of posts. Table 4 shows the chosen questions (thread titles) and the extracted keywords for each query.

3.2 Ground Truth

We manually find the ground truth of relevant results for each query, based on the analyzed user expectation as discussed in Section 1. We assume AND semantics among all the keywords in a query. To generate the ground truth for a query, we first define a relevant thread as a thread that contains all the query keywords. Consider the intensive human labor, we randomly choose 30 relevant threads for manual checking if a query involves more than 30 relevant threads. Since a patient is a semantic unit, we find the relevant patients whose associated information contains all the query keywords from the relevant threads. Then we consider the posts that are associated with such patients and contain at least one query keyword in the associated information as ground truth.

Table 4. Ten randomly chosen questions and keywords extracted from them

	Query questions (Keywords are underlined)
1	can **sz**[5] **type**s change?
2	**New Seizure**s...What does this mean?
3	Has your **temporal lobe epilepsy** become **worse** over **time**?
4	What is the **difference** in recordings between an **ambulatory EEG** and nonambulatory **EEG** (without **stimulus**)?
5	Anyone have a **child** with **Alternating Hemiplegia**?
6	Has anyone tried **Stiripentol** with their **kid**s?
7	**Growth** Spurt - **Breakthrough seizure**s?
8	**Vitamin**s to help with **aggression**???
9	**Seizure** due to **wean**ing?
10	**tonic** clonic after **flu virus**?

Table 5. Evaluation for ten randomly chosen queries

	Post-based			*Thread-based*			*Patient-based*		
Query	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*
1	0.978	0.379	0.547	0.509	1.0	0.674	0.764	0.698	0.73
2	0.973	0.379	0.545	0.477	1.0	0.646	0.702	0.695	0.698
3	1.0	0.059	0.111	0.378	1.0	0.548	0.923	0.706	0.8
4	1.0	0.333	0.5	0.5	1.0	0.667	1.0	1.0	1.0
5	1.0	0.286	0.444	0.636	1.0	0.778	1.0	0.714	0.833
6	1.0	0.5	0.667	0.5	1.0	0.667	1.0	0.5	0.667
7	1.0	0.063	0.118	0.087	1.0	0.16	0.8	0.75	0.774
8	1.0	0.429	0.6	0.28	1.0	0.438	1.0	0.714	0.833
9	1.0	0.534	0.696	0.349	1.0	0.518	0.716	0.658	0.686
10	1.0	0.3	0.462	0.455	1.0	0.625	0.6	0.3	0.4
Overall	0.995	**0.326**	**0.491**	**0.417**	1.0	**0.589**	**0.851**	**0.674**	**0.752**

[5] "sz" is identified as "seizure" using the acronym list in `http://epilepsyfoundation.ning.com/forum/topics/acronym-thread`.

3.3 Comparison Systems

We compare the ground truth with post-based search, thread-based search, and our approach, referred as patient-based search. Post-based search returns all the posts each containing all the query keywords. Thread-based search returns all the posts each containing at least one query keyword in a relevant thread. Our patient-based search returns all the posts each containing at least one query keyword associated with a relevant patient. Note that our approach shares the same intuition as the ground truth, but automatically identifies patients and automatically associate information to each patient. The quality of these automated processes has been evaluated.

3.4 Evaluation Metrics

We use standard evaluation metrics in information retrieval: precision (P), recall (R), and f-measure ($F1$). 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 ground truth. f-measure is defined as the harmonic mean of precision and recall: $F1 = \frac{2*P*R}{P+R}$.

3.5 Evaluation Results

The experimental results are shown in Table 5. It shows that post-based approach has almost perfect precision as in most cases keywords in the same post refer to the same patient and have close relationship, 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, but it has a very low precision. In contrast, our patient-based search has good precision and recall in general, and achieves a much higher f-measure than the other two approaches.

We also analyzed the major reasons that affect our system performance. First, some forum acronyms cannot be recognized, like "my DD" cannot be identified as "my daughter". Second, some patients cannot be identified by our system due to informal language used in a forum and the limited context. Third, some assumed reply relationships between posts are incorrect. We plan to leverage the method proposed in [9] to extract more accurate reply relationships between posts. Fourth, the performance of the current NLP tools, especially the co-reference resolution tool, is not perfect.

4 Conclusion and Future Work

To the best of our knowledge, this is the first work that makes patient-centered information extraction on healthcare forum. By building a database of patient information, we can process user search on the semantic units in the forum (patients) rather than the syntactic units (posts or threads) and thus achieve

high quality in information search. Our experimental evaluation verifies the effectiveness of our approach.

In future, besides addressing the several problems that we analyzed in experimental evaluation discussed earlier, we will also investigate the following issues to further improve our system. First, we will relax the AND semantics and develop a ranking model that ranks the results based on the relevance of the patients discussed in the results. Second, since obtaining a ground truth in this application is extremely labor-intensive, we will also investigate obtaining ground truth through crowdsourcing, where the challenge is how to design tasks for the crowd and how to consolidate their opinions to obtain ground truth.

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.

References

1. Stanford core NLP tools, `http://nlp.stanford.edu/software/corenlp.shtml`
2. Aronson, A.R.: Metamap: Mapping text to the UMLS metathesaurus (2006), `http://skr.nlm.nih.gov/papers/references/metamap06.pdf`
3. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research 12, 2493–2537 (2011)
4. Duan, H., Zhai, C.: Exploiting Thread Structures to Improve Smoothing of Language Models for Forum Post Retrieval. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 350–361. Springer, Heidelberg (2011)
5. Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D.: Stanford's multi-pass sieve coreference resolution system at the CoNLL 2011 shared task. In: Proceedings of ACL CoNLL 2011 Shared Task (2011)
6. Lindberg, D., Humphreys, B., McCray, A.: The unified medical language system. Methods of Inf. Med. 32(4), 281–291 (1993)
7. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: A corpus annotated with semantic roles. Computational Linguistics 31(1) (2005)
8. Seo, J., Croft, W.B., Smith, D.A.: Online community search using thread structure. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM) (2009)
9. Wang, H., Wang, C., Zhai, C., Han, J.: Learning online discussion structures by conditional random fields. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (2011)