Connecting the Dots: Knowledge Discovery in Online Healthcare Forums

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

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

ABSTRACT

Online healthcare forums provide a valuable platform for people to discover knowledge. However, existing approaches rely on the syntactic information units, such as a sentence, a post, or a thread, to bind different pieces of information in a forum. In this work, we propose to connect the pieces of information by a semantic information unit, patients. Specifically, we connect the information of diseases, symptoms, treatments, and effects to the patient who experiences them, and thus enable effective knowledge discovery.

Categories and Subject Descriptors

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

Keywords

Health Informatics; Knowledge Discovery; Healthcare Forums; Adverse Drug Reaction.

1. INTRODUCTION

Online healthcare forums are a popular platform for people to share their personal experience, participate in discussions, express their feelings, and to support each other. There are a lot of online healthcare forums available, such as PatientsLikeMe [1], WebMD [2], Healthboards message boards [3], MedHelp [4], and the Epilepsy forum [5]. The user population of such forums are rapidly growing. For instance, MedHelp currently has 13 million active monthly users. With highly valuable patient-contributed information and the ever increasing volume, such healthcare forums provide the potential for doctors and medical researchers to discover knowledge about various diseases, treatments, their effects and adverse reactions, and so on. For instance, adverse drug reaction (ADR) has become a leading cause of death in the U.S. [4]. While traditional small-scale patient surveys or voluntary report systems [1] are used to detect ADR, online healthcare forums may provide evidences in a much larger scale and in a timely fashion through the active participation of patients and caregivers. Furthermore, achieving "smart health and well-being" demands patients to take an active role in understanding their health status and in making informed decisions. Online healthcare forums enable patients and caregivers to perform searches regarding to specific questions.

Consider a user who wants to check other patients' experience of using Vitamin for alleviating aggression in order to gain more knowledge. She would issue a keyword query "Vitamin, aggression" on a healthcare forum. Many forums, such as WebMD and Patientslikeme, take an approach, referred as *post-based search* in this paper, that returns a post as a search result if it contains all the input query keywords. Consider a scenario where a caregiver describes her daughter has "aggression" and seeks for suggestions, and another experienced forum user replies and suggests her daughter to take "Vitamin" without explicitly quoting the word "aggression". While this would perfectly answer the user's query, it will be missed by the posted-based search as the query keywords appear in two posts instead of one. Therefore, post-based search may suffer low recall.

To improve the recall, another approach, named as *thread-based search* in this paper, and its variants are commonly used in many forums, such as the Healthboards message boards and the Epilepsy forum. It returns a thread or its link if all the posts in this thread collectively contain all the user input query keywords. Such an approach would be able to return the relevant result described earlier for the example query. However, suppose a user who suffers from seizures due to weaning wants to check other similar patients' experience. She would issue a keyword query "seizure, wean". Consider another thread discussing the effects of Vitamin B6 on an epilepsy patient. One post author mentions that her mother is taking Keppra to control her seizures while another caregiver mentions her son has weaned off Keppra since it causes anger. Although both of them have benefited from Vitamin B6, nobody has "seizure" due to "weaning". This thread is thus not relevant to the user's search intention. Such an irrelevant thread will be returned by the thread-based search as a query result, suffering low precision.

To understand the root cause of the problem, let us first understand the semantics of the user query. By issuing a query "Vitamin, aggression" or "seizure, wean", the user would like to find out the effects of Vitamin to aggression or the seizures caused by weaning happened on a patient. In other words, given multiple query keywords, *the dots*, a user would like to *connect* them by a patient whose experiences relate to all the query keywords. However, such

[1] http://www.patientslikeme.com
[2] http://exchanges.webmd.com
[3] http://www.healthboards.com/boards
[4] http://www.medhelp.org/
[5] http://epilepsyfoundation.ning.com/forum

ICEC'14 August 05 - 06 2014, Philadelphia, PA, USA.
 ACM 978-1-4503-2618-6/14/08...$15.00.
http://dx.doi.org/10.1145/2617848.2617864.

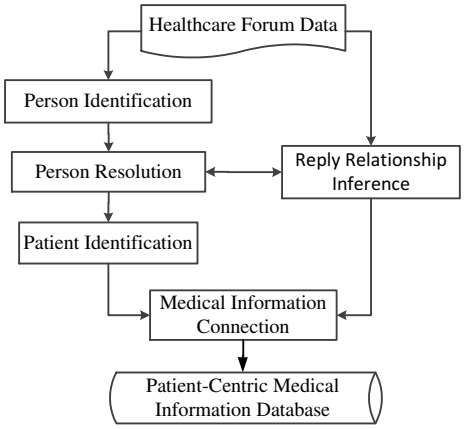

Figure 1: The system overview.

connection is not readily available in the data. Existing approaches choose to connect the dots by syntactic information unit, such as a post or a thread. Unfortunately, syntactic information units (e.g. a sentence, a post or a thread) used by existing approaches are typically misaligned with semantic information unit (a patient) that a user relies on, resulting in poor search quality.

This problem not only exists in patients' information search, but also occurs in researchers' knowledge discovery on online healthcare forums. Recently, work has been performed on ADR discovery in online healthcare forums through co-occurrence analysis [8, 10, 15]. It assumes that if two medical phrases, such as a drug "Depakote" and an ADR "fever", co-occur in the same syntactic unit (e.g. a sentence, a post or a thread) frequently, then "Depakote" is considered to cause "fever". However, such occurrences of "Depakote" and "fever" may not refer to the same patient.

In this work, we propose to connect the dots (i.e. the pieces of information) by a semantic information unit (i.e. a patient) for effective knowledge discovery in online healthcare forums. In particular, we connect the information of conditions, symptoms, diseases, treatments, and effects to the patient who experiences them. With the well-connected information, we can support patients to perform effective information search and support researchers to perform knowledge discovery.

Related Work: There is a growing interest in knowledge discovery from online healthcare social media. [6] focuses on information integration from multiple online communities to obtain comprehensive information. [8, 10, 15] study the problem of detecting ADR through co-occurrence analysis from social media with different syntactic document units. In contrast, we propose to connect information in healthcare forums via a semantic unit, a patient. Opinion mining and sentiment analysis are also used for knowledge discovery in online healthcare forums [5, 9]. [9] discovers patient drug outcomes by clustering the topics and opinions in online health forums. [5] compares the effectiveness of different treatments with sentiment analysis. The user's demographic information is also used in [5] to differentiate the treatment effectiveness for different populations. Different from our work, they do not connect the medical information within a patient semantic unit. Some empirical studies of our work in terms of search are presented in [12, 13].

2. SYSTEM OVERVIEW

We developed PCMIE, a Patient-Centric Medical Information Extraction system, to connect the medical information within a patient semantic unit. The PCMIE system takes the forum data as input and generates a patient-centric medical information database as output. Fig. 1 shows the system architecture. To connect the pieces of information that are related to the same patient, we need to identify person mentions in the posts in *Person Identification* module, then the *Person Resolution* module connects all the person mentions that refer to the same person. While post reply relationships are useful for person resolution, person reference relationships are also used to infer the unknown post reply relationships for better medical information connection, as achieved in the *Reply Relationship Inference* module. The *Patient Identification* module then identifies if a person is a patient based on the context of all the person mentions referring to the same person. At last, *Medical Information Connection* module connects the medical information pieces with the relevant patient mentions. Next we discuss each module.

The *Person Identification* module first identifies all the person mentions in a post using Stanford NLP [2] and MetaMap tool [3]. A person mentioned in a healthcare forum may serve different roles, such as a patient, a caregiver, or a doctor. For example, from the sentence "My daughter was diagnosed last Friday", "My daughter" is labeled as a patient by semantic role labeling (SRL) [7] with Propbank annotation [14]. However, it may not always be straightforward to determine whether a mentioned person is a patient. Consider another sentence: "She is 5 years old". We cannot identify if "She" is a patient. But we observe that if "she" refers to the same person as 'My daughter", then we know "she" is a patient mention. Thus, in order to effectively identify whether a person is a patient, we need to perform person resolution to find all mentions of the same person and thus obtain a big context for judgement.

In the *Person Resolution* component, we cluster person mentions in posts such that all the mentions in the same cluster refer to the same person. There are two types of person resolution: intra-post and inter-post person resolution. For intra-post person resolution, we use the Stanford co-reference resolution system [2] to cluster all the person mentions within each post. Since co-reference resolution is more general than person resolution, we can easily extract the person resolution results from the co-reference resolution results.

While existing work is focused on the resolution within a single document (a post), an open challenge in healthcare forums is that the same person can be mentioned in different posts. To address that, we have developed techniques for inter-post person resolution. We first merge the person clusters if the person mentions serve the same role related to the same post author. For example, if the same author mentioned "my daughter" in two different posts, the two person clusters containing these two mentions are merged. We also propose to utilize post reply relationships to merge clusters across posts. For example, if post A mentions "my daughter", and post B replies to post A and mentions "your daughter", then we know these two mentions refer to the same person and merge the two corresponding clusters.

As we can see from the previous example, the reply relationships between posts provide valuable information for inter-post person resolution. In fact, it can also be used for information connection in the *Medical Information Connection* module. However, such reply relationships may be unknown since some forum users do not specify the reply relationship explicitly. We develop effective techniques to infer the unknown reply relationships in the *Reply Relationship Inference* module [11]. Among diverse features we consider, one interesting observation is that: person reference relationships can be used to infer unknown reply relationships.

After grouping person mentions into clusters, in the *Patient Identification* module, we identify if a person is a patient based on all the information in the cluster. Semantic role labeling (SRL) [7] with Propbank annotation [14] and some patterns are used for pa-

tient identification. For example, in the sentence "My daughter was diagnosed last Friday", "My daughter" is labeled as a patient, as well as all the person mentions in the same cluster.

At last, in the *Medical Information Connection* module, we connect the pieces of medical information, such as diseases, symptoms, treatments, effects, with the closest patient mention before them, and store all these connections in a patient-centric database. Note that a piece of information, like the word "Vitamin" that appears in a replying post in the earlier example, can be connected with a patient mentioned in a parent post.

3. APPLICATIONS

With the generated patient-centric medical information database that connects medical information with the corresponding patients, we can provide more effective search on online healthcare forums [12]. Consider again the example query "Vitamin, aggression" or "seizure, wean", PCMIE finds the patient whose information contains both query keywords, even if the information appears in different posts. On the other hand, if the two keywords are not connected with the same patient, even if they appear in the same post, they will not be considered relevant to the query.

With large healthcare forum data available, we can perform effective co-occurrence analysis for ADR discovery. We first identify all the drug names and their alias, of which we want to discover their adverse reactions. A list of potential ADR phrases will be used to identify all their occurrences in the data. Then we use PCMIE to connect the drug names with the identified ADR phrases. Specifically, if a drug name and an ADR phrase co-occur in the same patient's information, then we consider the ADR is caused by that drug. By enforcing each co-occurrence refers to the same patient, we expect to achieve high accuracy in ADR knowledge discovery.

4. CONCLUSIONS AND FUTURE WORK

In this work, we propose to connect pieces of medical information, such as diseases, symptoms, treatments and effects, in online healthcare forums by their semantic information units, patients, for effective knowledge discovery. Toward this goal, we present our PCMIE system that performs natural language processing and data mining to generate a patient-centric database. Then we discuss two applications of using this system: keyword search and co-occurrence analysis for ADR discovery. We have applied the PCMIE system for keyword search on the Epilepsy forum data and obtained promising results [12]. We are currently applying it for co-occurrence analysis based ADR discovery on large-scale data. A unique challenge is that the connection type matters in this application. For example, after we identify that "Depakote" and "fever" are strongly connected based on their associated patients, the connection type is unclear: is drug "Depakote" used to treat "fever" or does "Depakote" cause "fever"? In other words, it is hard to determine whether fever is a symptom or an adverse effect, since there is a significant overlap between their vocabulary. We plan to classify the connection types between two connected information pieces that are associated with the same patient in order to improve the accuracy of knowledge discovery.

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.

5. REFERENCES

[1] FDA's Adverse Drug Event Reporting System (FAERS). http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm.

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

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

[4] B. W. Chee, R. Berlin, and B. Schatz. Predicting adverse drug events from personal health messages. *AMIA Annual Symposium Proceedings*, 2011.

[5] J. H. D. Cho, V. Q. Z. Liao, Y. Jiang, and B. R. Schatz. Aggregating personal health messages for scalable comparative effectiveness research. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, 2013.

[6] 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.

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

[8] B. A. et.al. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. *J. Biomed. Inf.*, 44, 2011.

[9] Y. Jiang, Q. V. Liao, Q. Cheng, R. B. Berlin, and B. R. Schatz. Designing and evaluating a clustering system for organizing and integrating patient drug outcomes in personal health messages. In *AMIA Annu Symp Proc*, 2012.

[10] X. Liu and H. Chen. Azdrugminer: An information extraction system for mining patient-reported adverse drug events in online patient forums. In *Proceedings of International Conference for Smart Health*, 2013.

[11] 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.

[12] 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.

[13] Y. Liu and Y. Chen. Patient-centric, multi-role, and multi-dimension information exploration on online healthcare forums. In *Proceedings of PIKM 2013*, 2013.

[14] M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1), 2005.

[15] C. C. Yang, H. Yang, and L. Jiang. Postmarketing drug safety surveillance using publicly available health consumer contributed content in social media. *ACM Transactions on Management Information Systems*, 5, 2014.