Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Exploring methods for identifying related patient safety events using structured and unstructured data

Allan Fong *, A. Zachary Hettinger, Raj M. Ratwani

MedStar Institute for Innovation – National Center for Human Factors in Healthcare, 3007 Tilden St. NW, Suite 7M, Washington, DC 20008, USA

ARTICLE INFO

Article history:
Received 31 March 2015
Revised 2 July 2015
Accepted 10 September 2015
Available online xxxx

Keywords:
Patient safety event reports
Topic modeling
Unstructured data
Structured data
Search
Relevance

ABSTRACT

Most healthcare systems have implemented patient safety event reporting systems to identify safety hazards. Searching the safety event data to find related patient safety reports and identify trends is challenging given the complexity and quantity of these reports. Structured data elements selected by the event reporter may be inaccurate and the free-text narrative descriptions are difficult to analyze. In this paper we present and explore methods for utilizing both the unstructured free-text and structured data elements in safety event reports to identify and rank similar events. We evaluate the results of three different free-text search methods, including a unique topic modeling adaptation, and structured element weights, using a patient fall use case. The various search techniques and weight combinations tended to prioritize different aspects of the event reports leading to different search and ranking results. These search and prioritization methods have the potential to greatly improve patient safety officers, and other healthcare workers, understanding of which safety event reports are related.

1. Introduction

In an effort to improve safety most healthcare systems have a patient safety reporting system (PSRS) in place [1,2]. These systems provide a method for staff, including physicians, nurses, and technicians, to report on safety events in their environment ranging from near misses, where harm almost reaches a patient, to serious safety events, where a patient is harmed [3]. The Institute of Medicine has strongly recommended the use of these systems to identify why patients are harmed by medical errors, and several states require the use of a PSRS [3].

Patient safety event (PSE) reports generally contain structured information such as the time and site of occurrence, role of the participants (physician, nurse, technician, etc.), patient demographic and clinical attributes, as well as a classification of the severity and type of event. In addition to the structured data elements the safety event reports also include an unstructured free-text field in which the reporter can provide a narrative describing the patient safety event in greater detail [4]. Patient safety reporting systems can grow to contain tens of thousands to hundreds of thousands of events. Patient safety organizations, which act as safe harbors that allow providers to share PSE data without the liability risk, serve to combine PSE reports from multiple providers resulting in even larger safety event databases.

If the data from patient safety reporting systems can be analyzed effectively the databases of reports hold tremendous promise for improving patient safety [5,6]. The data can be used to identify important patterns or trends of events that can then be remedied by intervening to remove or mitigate potential safety threats. However, to realize the promise of patient safety event reporting systems, efficient and effective analysis methods need to be developed to allow for a deeper understanding of the data that can then lead to action to improve safety. The challenge is that patient safety data are incredibly complex with both structured and unstructured data elements.

While analyzing the structured data may be relatively straightforward these data only provide a partial understanding of the safety event and many events actually span multiple pre-defined categories, such as with general event type (e.g. medication, falls, miscellaneous, diagnosis/imaging) and specific event type assignments. Given that many event reporting systems only allow the selection of one general event and specific event type category the structured data may not accurately reflect the context of the safety event. Furthermore, most safety reporting systems do not provide a formal definition of the event type category. As a result, the reporter must select a category based on their own knowledge and intuition, hence the ambiguity that can sometimes arise from these categories. For a more complete understanding of the event,

* Corresponding author.
E-mail address: allan.fong@medicalhfe.org (A. Fong).

http://dx.doi.org/10.1016/j.jbi.2015.09.011
1532-0464/

the unstructured free-text narratives, which often contain a rich description of the patient safety event from the perspective of the reporter, must be analyzed in combination with the structured data more efficiently and effectively.

A very important and practical use of the PSE data is to understand whether a recently reported event is part of a larger trend that should be of immediate concern and warrant the allocation of limited resources to address the safety hazard or whether the recent event is an anomaly and can be prioritized accordingly [5]. In this paper we focus on developing PSE report analysis techniques to address this very pressing problem. To understand how a new event aligns with previous events one must to be able to identify other PSE reports in the event report database that are similar. Although most patient safety reporting systems provide the ability to query by structured fields and match keywords in the unstructured narratives, this process remains labor intensive and challenging because of the large number of irrelevant reports that are returned from these searches.

In this paper, we present and explore methods for searching large databases of PSE data to identify and rank similar event reports using both unstructured and structured data. This approach consists of using natural language process (NLP) techniques to search free-text and a structured element weighting scheme to prioritize the search results. NLP leverages the power of computers to process and make sense of large amounts of text. There are several NLP methods and strategies developed for search and retrieval tasks. For example, identifying important words in reports and document distant metrics can be used to help find, match, and rank documents by their similarities [7–9]. In addition, methods, such as topic modeling, have been widely used to identify latent themes or topics in documents [10]. Reports that discuss similar topics would have similar topic probability profiles. Furthermore, previous work has used NLP approaches to categorize and identify health information safety events and extreme risk events in free-text reports [11,12]. NLP techniques have also been used to identify safety events from clinical documentations [13,14]. In addition to analyzing free-text, techniques such as structural topic modeling and labeled Latent Dirichlet Allocation (LDA) have been developed to take into consideration structured data, or meta data, into topic models [15–17]. However, these studies did not focus on, and evaluate, the utility of considering various structured data elements on search methods.

The work presented in this paper builds on these previous research efforts by utilizing the additional structured data elements in the PSE reports to search for and rank related events. While an analysis of the unstructured data alone has been a useful approach, leveraging the structured data elements in combination with a NLP approach may allow for the improved identification of similar PSE reports. However, combining the structured and unstructured data is not always intuitive and it can be difficult to interpret how the inclusion of different structured elements impacts search results. These challenges are further exacerbated by noisy structured data, particularly prevalent in self-reported PSE data. In our approach, we propose a more intuitive and less complex approach to incorporate and evaluate the effects of considering structured elements with the unstructured free-text data to enhance the search for similar PSE reports.

2. Methods

Our approach consists of leveraging both the free-text and structured elements in PSE reports to identify and rank related PSE reports. We first discuss three different NLP search methods (topic models, unigrams, and bigrams) and their application to the unstructured free-text in each report. Each method takes a single PSE report, the base report, and ranks other PSE reports according to their similarity to the base report. We present a unique approach that leverages topic modeling results to find related PSE reports and compare these results with more standard unigram and bigram search techniques. We then discuss how the structured element weights are incorporated with the ranking results.

2.1. Data

PSE reports were collected, through self-report, over a two year period (January 2013 to January 2015) from a multi-hospital healthcare system in the mid-Atlantic region of the United States. A total of 49,859 reports were collected during this time from the PSRS. Each report has both structured and unstructured fields. For this analysis, we focused on the general event type (GET) and specific event type (SET) structured elements (as they provide the greatest insight about the reports compared to other structured elements) and the unstructured free-text brief factual description of the event. There are 21 GETs, the most common being "medication/fluid," "labs," "falls" and "miscellaneous," Table 1. Furthermore, each GET category is comprised of several unique SET categories. For example, the "falls" GET category has "from bed," "while ambulating," and "from chair" SET categories while the "medication/fluid" GET has "adverse drug reaction," "duplicate therapy," and "wrong patient" SET categories. Every PSE report is assigned a single GET and a SET category by the reporter of the event.

2.2. Free-text search methods

We present three free-text search methods below. Punctuations, numbers, common stop words were removed from the free-text and words were stemmed prior to analysis. After this preprocessing, the medium number of terms (words) for a report was 27 with a standard deviation of 33.

2.2.1. Topic model

Our first method adapted topic modeling techniques to evaluate the similarity or relevance between reports. We first used the term frequency – inverse document frequency (tf-idf) statistic to identify important words in each report [8]. We then used this subset of words as inputs to our LDA topic model [10]. Reports were evaluated based on their topic probability distribution distance from a base report.

While LDA is a popular and commonly used topic modeling technique, it is limited, particularly if the underlying topics are not well-separated [18]. This is more likely to occur when the informational content in documents is noisy. Unfortunately, PSE reports greatly vary in complexity and length; some reports are brief sentences while others are long detailed narratives. To address this limitation, we used tf-idf to first identify the

Table 1
Top 10 GET by percent of total reports.

General event type categories	Percent
Medication/fluid	17
Lab/specimen	15
Fall	12
Miscellaneous	10
Blood bank	7
Skin/tissue	5
Diagnosis/treatment	5
Patient ID/documentation/consent	5
Surgery/procedure	4
Lines/tubes/drain	4

Please cite this article in press as: A. Fong et al., Exploring methods for identifying related patient safety events using structured and unstructured data, J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.09.011