A Novel Approach for Adverse Events Detection

Jia Zhu[1], **Yuzhi Liang**[2], **Changqin Huang***[1], **Min Yang**[2], **Jing Xiao**[1], **Yong Tang**[1]
[1]School of Computer Science, South China Normal University, Guangzhou, China
*cqhuang@scnu.edu.cn
[2]School of Computer Science, The University of Hong Kong, Hong Kong, China

Abstract

Adverse events detection is critical in many fields, e.g., adverse drug event (ADE) detection in medical field. ADE is an unexpected and harmful consequence of drug usage. Many researchers found that identifying the correlation between the use of drugs and adverse events from biomedical literature can contribute a lot to drug safety monitoring. In this paper, we propose a novel approach based on biomedical literature to detect ADE. We first construct a graph using candidate ADE extracted from biomedical literature, and then propose a method to select critical vertices from the graph as core vertices with a clustering algorithm to group these core vertices to build subgraphs. Lastly, the correlations between drugs and events are calculated based on the subgraphs for ADE detection. Experiments show that our approach is highly feasible.

1 Introduction

Adverse events detection, such as adverse drug event (ADE) detection, is very important in drug safety assessment. It encompasses all scientific and data gathering activities relating to the detection, assessment, and understanding of adverse events of medical products through the product life cycle [Saless, 2005]. ADE refers to any unfortunate medical and health events that occur during the course of drug treatment, and this event does not necessarily have a causal relationship with drug therapy [Liu *et al.*, 2016; Cai *et al.*, 2017]. ADE might be caused by several drugs interacting with each other when administered concomitantly. Though drug standards are guided for the production and usage of drugs, it will still lead to the occurrence of ADE even if we obey the criteria drug standards.

As we all know, traditionally drug safety monitoring relies on data from spontaneous reporting systems (SRS), such as the US Food and Drug Administration' (FDA) Adverse Event Reporting System (FAERS), which contains reports of suspected ADE submitted by health care providers, manufactures, and patients. However, FAERS data do have some limitations. For example, FDA does not receive reports for every adverse event or medication error that occurs in a product. Therefore, FAERS data is not sufficient to calculate the incidence of an adverse event or medication error in the U.S. population.

Since there is no sufficient medical data or authoritative evidence to identify the correlations between drugs and adverse events quickly, it is challenging but extremely necessary to detect ADE effectively using other methods. In this paper, we focus on using the data extracted from biomedical literature according to [Winnenburg and Shah, 2016]. We propose a novel graph-based approach called G-ADE to detect ADE. In G-ADE, we first construct a graph using candidate ADE extracted from biomedical literature. We then propose a novel method to select important vertices from the graph as core vertices, and design an algorithm using these core vertices for clustering in order to build a set of subgraphs. This step is our main contribution in this paper because finding core vertices in a graph has been proved extremely useful for later processing [Yang *et al.*, 2016; Min *et al.*, 2009; Yang *et al.*, 2017; Yang and Chow, 2014]. Lastly, the correlation between each drug-event pair is calculated based on the subgraphs we constructed in the previous step to identify ADE. We have also perform a few experiments to validate our work using a gold standard of drug-adverse event correlations spanning 159 drugs and four events.

The rest of this paper is organized as follows. In Section 2, we describe details of G-ADE including a core vertices selection algorithm with subgraphs generation. In Section 3, we describe our experiments with evaluation methods and result analysis, and compare G-ADE with other methods. We also conclude and discuss this study in Section 4.

2 Proposed Approach

The overview of G-ADE is described in Figure 1. The process can be summarized as follows. Firstly, we extract all associated pairs between all adverse events and drugs from MEDLINE according to [Winnenburg and Shah, 2016], and use these pairs as basic knowledge to construct a fully connected drug-event graph. These pairs represent a list of potential ADE to be validated. Secondly, we propose a core vertices selection algorithm to select core vertices in the drug-event graph we built. Thirdly, we adopt an algorithm to generate a set of subgraphs based on the core vertices obtained in the previous step based on the idea from [Min *et al.*, 2009] for the purpose of key information extraction. Finally, we connect all the subgraphs into a graph that in fact is a smaller size

of the original graph, and compute the similarity of all drug-event pairs based on this new graph. If the drug-event pair's similarity is higher than the threshold we set up, we will recognize it as ADE. We will introduce the details of each step in the following sections.

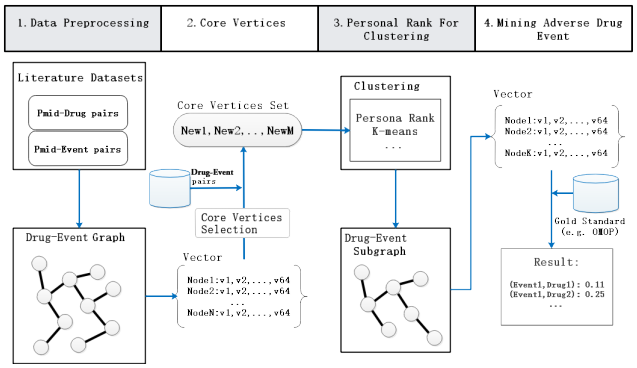

Figure 1: Overview of G-ADE

2.1 Core Vertices Selection

As we described earlier, the key is to select critical vertices as core vertices in order to construct subgraphs. We propose a method that computes each vertex's importance based on vectors generated by DeepWalk [Perozzi *et al.*, 2014]. DeepWalk has been proved successfully in social networks and graph analysis. It learns the latent representations by modeling a stream of short random walk, and then encodes it in a continuous vector space with low dimensions. In our purposed method, we apply DeepWalk to the drug-event graph to get a 64-dimensions vectors for each vertex. Let $G = (V, E)$ to be the drug-event graph, $v \in V$, which represents a drug or an event. H is the set of neighbor vertices of v, $H_i \in H$, n is the number of the neighbor vertices. We have Equation (1) to compute the score of importance for each vertex. The core vertices selection algorithm to adopt the score is described in Algorithm 1.

$$score = \frac{\sum_{i=1}^{n} Dist_{v,H_i}}{n} \tag{1}$$

2.2 Personal Rank for Clustering

Once we have a list of core vertices, we adopt Personal Rank(PR) [Haveliwala, 2003] algorithm to generate associated subgraphs. PR is a graph clustering algorithm based on random walk [Haveliwala, 2003], which performs well in recommendation system and social network [Li *et al.*, 2012; Shen *et al.*, 2016]. The basic idea of PR is similar to PageRank [Bianchini *et al.*, 2005]. Firstly, the algorithm computes the score of importance of each vertex in the network, and then sorts each node according to the score of the importance. Eventually, the algorithm outputs Top-N vertices. The score of importance is defined in Equation (2):

$$PR_i = (1-\alpha)r_i + \alpha \sum_{j \in in_i} \frac{PR_j}{|out_i|}, r_i = \begin{cases} 1 & i = v_{core} \\ 0 & i \neq v_{core}, \end{cases} \tag{2}$$

Algorithm 1 Core Vertices Selection Algorithm

Require: $G = (V, E)$, a drug-event graph, t, a threshold used to filter
Ensure: $List$, a sorted list of core vertices
1: learns the $vectors$ of the V, $vectors$ are the DeepWalk Vectors
2: **for** each $v \in V$ **do**
3: **for** each $H_i \in H$ **do**
4: compute $Dist_{v,H_i}$, the distance between v and H_i using $vectors$
5: **end for**
6: compute v'score using Equation (1)
7: **end for**
8: sorted v by its score, and put v which score is higher than t into $List$

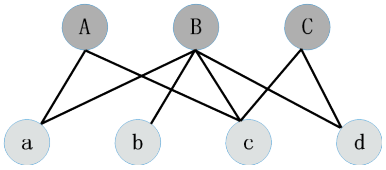

Figure 2: Example of Personal Rank

where PR_i represents the score of importance of vertex i, α is the probability of walking, out_i represents the out degree of vertex i and in_j represents the in degree of vertex j, v_{core} represents the core vertex. As shown in Figure 2, we set PR of core vertices(A,B,C) to 1 and other vertices(a,b,c,d) to 0 initially. Therefore, we have $PR(A) = PR(B) = PR(C) = 1$, and $PR(a) = PR(b) = PR(c) = PR(d) = 0$ in the beginning, then we starts to walk from the vertex with $PR \neq 0$. The probability of walking is α, correspondingly, the probability of stopping is $1 - \alpha$. For example, the probability of walking from A to a or c is 0.5 as a and c share the importance of A, that is $PR(a) = PR(c) = PR(A) * 0.5$. The process then continues to walk with probability of α or stop and go back to A with probability of $1 - \alpha$. The process will keep running until the PR of each vertex is stable. We select up to top-100 highest PR vertices that attached to each core vertex to construct subgraphs. In other words, we will get a set of clusters/subgraphs in the end of this process.

2.3 Mining Adverse Drug Events

As now we have a set of clusters(subgraphs), we next want to transform all the clusters into a graph for further processing. We define S_g is the set of clusters we got in the previous step, and $G = (V, E)$ is the original graph. If $G_a = (V_a, E_a)$ and $G_b = (V_b, E_b)$ are two clusters in S_g, we put G_a and G_b into a new graph if there is an edge between v_a and v_b in G, and $v_a \in V_a$, $v_b \in V_b$. We continue this process until all possible clusters are checked, and a new graph is generated eventually. We then calculate the similarity between drug vertices and event vertices in this new graph according to their associated vectors generated by DeepWalk. Note that the reason we run DeepWalk again on the graph to generate new vector for each vertex because the

structure of the graph is changed. If two vertices' associated vectors are very similar, then means these two vertices has very closed information in the graph [Perozzi *et al.*, 2014; Cunchao Tu, 2016; Yang and Liu, 2015; Zhu *et al.*, 2016; 2012]. In other words, this drug-event pair has higher probability to be an ADE. Therefore, according to the similarity of each pair, we can mine ADE from the graph if the similarity of drug-event pair is higher than a certain threshold we set.

3 Experiments

In this section, we will introduce the evaluation we have done on OMOP standard data set [Ryan *et al.*, 2013]. The experimental results show evidence of significant improvement of our proposed approach over baseline methods.

3.1 Datasets

Literature Data Set

The data set used in this paper can be downloaded from the website[1]. It consists of candidate ADE pairs extracted from MeSH term indexes of all 366k articles in MEDLINE that are indexed with certain combinations of MeSH terms and qualifiers according to [Winnenburg and Shah, 2016]. The creation of this data set is described in detail in [Winnenburg *et al.*, 2015]. We extract PMID-Drug pairs (paper id and drug) and PMID-Event pairs (paper id and event) from data set. The detail data set is shown in Table 1.

Name	Number
Pmid	366k
Drug	3416
Event	1602
Pmid - Drug pairs	418k
Pmid - Event pairs	552k

Table 1: Experimental data set

Gold Standard Data Set

We evaluate the experimental performance against the drug safety reference set established by the observational medical outcome partnership (OMOP) [Ryan *et al.*, 2013]. This set contains 399 drug-outcome pairs, covering 183 drugs from several drug classes and four significant and actively monitored adverse event outcomes: acute myocardial infarction, acute renal failure, acute liver injury, and upper gastrointestinal bleeding. Here we only select the positive part that indicates ADE as the ground truth to validate our approach. Note that we have removed a few drugs that cannot be found in our graph. The data set is shown in Table 2.

[1]ftp://nlmpubs.nlm.nih.gov/online/mesh/2015/meshtrees/

Aggregation Outcome	Drugs	
	Positive	Negative
Acute kidney injury	23	59
Acute liver injury	79	33
Acute myocardial infarction	33	61
GI bleed	24	62
Total	159	215

Table 2: OMOP data set

3.2 Baseline Methods

In order to evaluate the feasibility of G-ADE and the core vertices selection algorithm, we use the following two methods(BG and CVTn) as the baseline methods.

* **Basic Graph(BG):** In this approach, We firstly construct a graph using all candidate ADE extracted from biomedical literature as described in Section 3.1. We obtain the corresponding vector of event and drug through the DeepWalk algorithm. Finally, their correlations are calculated by cosine similarity, namely, basic graph(BG).
* **Core Vertices from Top n Nodes(CVTn):** In this method, we construct the graph using the same measure as BG. We then simply select top N nodes as the core vertices according to the degree of the node. We rebuild new subgraphs using the core vertices with a clustering algorithm. Lastly, the correlation/similarity between the drug and the event is calculated based on the subgraphs. We name it as CVTn.

3.3 Evaluation Standard

In the experiment, we calculate the similarity between event and drug to according to the associated vector of each vertex. The cosine similarity measure [On, 2008] between two vectors is used that calculates the cosine of the angle between them, and the cosine similarity formula is defined as:

$$Similarity_{Event,Drug} = \frac{\overrightarrow{V_{Event}} \times \overrightarrow{V_{Drug}}}{||\overrightarrow{V_{Event}}|| \times ||\overrightarrow{V_{Drug}}||} \quad (3)$$

The higher the similarity is, the more this event relates to the drug. At the same time, we verify the accuracy of the standard data set by different thresholds. If the similarity between drugs and events is higher than the threshold, we think they are relevant. That is, the drug has an adverse effect on the event. Therefore, we can calculate the predication accuracy as: $Accuracy = N/M$, where N is the number of the drug-even pair matched the ADE in OMOP and M is the number of ADE in OMOP.

Thr	BG	CVTn		G-ADE	
		PR	K-means	PR	K-means
0.1	39.6%	42.1%	42.8%	**50.3%**	48.4%
0.2	74.8%	75.4%	76.1%	**79.2%**	78.0%
0.3	85.5%	86.8%	86.1%	**91.1%**	89.9%

Table 3: Prediction Accuracy in the OMOB data set.

3.4 Evaluation Results

The overall performance of G-ADE and two baselines(BG,CVTn) methods in terms of accuracy at different thresholds using the OMOP data set as reference is summarized in Table 3. Note that since the third step for clustering of CVTn and G-ADE can both be replaced by other clustering algorithms, e.g., K-means [Kanungo *et al.*, 2002]. Therefore, we have also provided the evaluation results produced by K-means. In our experiments, we set $K = 10$ as this setting can achieve the best performance on our dataset.

From the Table 3, we can clearly see that G-ADE outperforms the BG and CVTn methods with different clustering algorithms (Personal Rank and K-means) at different thresholds. Obviously, the prediction accuracy is improved for all approaches with higher threshold. On the other hand, though CVTn method performs better than the BG method but not as good as G-ADE with both clustering algorithms, which demonstrates that the core vertices selection plays an important role to improve prediction accuracy. Overall, our proposed approach based on core vertices selection and PG clustering for the detection of adverse drug events is feasible as 91.1% prediction accuracy is achieved. Therefore, we can conclude that through the core vertices selection and personal rank clustering algorithm, the accuracy of adverse drug event detection is effectively improved because the robust performance is shown in our experiments.

4 Conclusions

Adverse drug event (ADE), defined as adverse patient outcomes caused by medications, is a common issue but difficult to detect. In this paper, We propose an approach called G-ADE based on biomedical literature to detect ADE. We first construct a graph using candidate ADE extracted from biomedical literature. We then propose a core vertices selection algorithm to select important vertices from the graph as core vertices, and design a PR algorithm using the core vertices for clustering to build subgraphs. Last but not least, the correlation between the drug and the event is calculated using the vector generated by DeepWalk based on the subgraphs. If the correlation value is sufficiently high, we take this pair as ADE. Our experimental results show that G-ADE performs better than two baseline methods in a few scenarios. In the future, we will pay more attention to build subgraphs on ensemble different clustering methods to further improve the performance and attempt to gain data from social networking as this is where a lot of patients may talk about uncommon side effects from medication and hence could lead to interesting clinical results.

5 Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (61370229), the Natural Science Foundation of Guangdong Province, China(2015A030310509), the Public Research and Capacity Building in Guangdong Province, China(2016A030303055), the S&T Projects of Guangdong Province, China(2015A030401087, 2016B030305004, 2016B010109008), and the S&T Project of Guangzhou Municipality(201604010054).

References

[Bianchini *et al.*, 2005] Monica Bianchini, Marco Gori, and Franco Scarselli. Inside pagerank. *Acm Transactions on Internet Technology*, 5(1):92–128, 2005.

[Cai *et al.*, 2017] Ruichu Cai, Mei Liu, Yong Hu, Brittany L. Melton, Michael E. Matheny, Hua Xu, Lian Duan, and Lemuel R. Waitman. Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports. *Artificial Intelligence in Medicine*, 76:7–15, 2017.

[Cunchao Tu, 2016] Zhiyuan Liu Maosong Sun Cunchao Tu, Weicheng Zhang. Max-margin deepwalk: Discriminative learning of network representation. In *IJCAI*, pages 1–7, 2016.

[Haveliwala, 2003] Taher H Haveliwala. Topic-sensitive pagerank. In *International Conference on World Wide Web*, pages 517–526, 2003.

[Kanungo *et al.*, 2002] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Analysis and Machine Intelligence.*, 24:881–892, 2002.

[Li *et al.*, 2012] Xin Li, Xin Su, and Mengyue Wang. Social network-based recommendation:a graph random walk kernel approach. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 409–410, 2012.

[Liu *et al.*, 2016] Jing Liu, Songzheng Zhao, and Xiaodi Zhang. An ensemble method for extracting adverse drug events from social media. *Artificial Intelligence in Medicine*, 70:62–76, 2016.

[Min *et al.*, 2009] Wu Min, Xiaoli Li, Chee Keong Kwoh, and See Kiong Ng. A core-attachment based method to detect protein complexes in ppi networks. *BMC Bioinformatics*, 10(1):169, 2009.

[On, 2008] Byung Won On. Social network analysis on name disambiguation and more. In *International Conference on Convergence and Hybrid Information Technology*, pages 1081–1088, 2008.

[Perozzi *et al.*, 2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: online learning of social representations. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710, 2014.

[Ryan *et al.*, 2013] P. B. Ryan, M. J. Schuemie, E Welebob, J Duke, S Valentine, and A. G. Hartzema. Defining a reference set to support methodological research in drug safety. *Drug Safety*, 36(1):33–47, 2013.

[Saless, 2005] Fatieh Saless. Review of fda guidance on risk management: Good pharmacovigilance practices and pharmacoepidemiologic assessment. *Genetic Engineering News*, 25(18):14–0, 2005.

[Shen *et al.*, 2016] B Shen, B Hu, and H Zhang. Method for the analysis of the preferences of network users. *Networks Iet*, 5(1):8–12, 2016.

[Winnenburg and Shah, 2016] R. Winnenburg and NH Shah. Generalized enrichment analysis improves the detection of adverse drug events from the biomedical literature. *BMC Bioinformatics*, 17(1):1–17, 2016.

[Winnenburg *et al.*, 2015] Rainer Winnenburg, Alfred Sorbello, Anna Ripple, Rave Harpaz, Joseph Tonning, Ana Szarfman, Henry Francis, and Olivier Bodenreider. Leveraging medline indexing for pharmacovigilance - inherent limitations and mitigation strategies. *Journal of Biomedical Informatics*, 57(C):425, 2015.

[Yang and Chow, 2014] Min Yang and Kam-Pui Chow. Authorship attribution for forensic investigation with thousands of authors. In *IFIP International Information Security Conference*, pages 339–350. Springer Berlin Heidelberg, 2014.

[Yang and Liu, 2015] Cheng Yang and Zhiyuan Liu. Comprehend deepwalk as matrix factorization. *Computer Science*, pages 1–7, 2015.

[Yang *et al.*, 2016] Min Yang, Jincheng Mei, Fei Xu, Wenting Tu, and Ziyu Lu. Discovering author interest evolution in topic modeling. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 801–804. ACM, 2016.

[Yang *et al.*, 2017] Min Yang, Dingju Zhu, Yong Tang, and Jingxuan Wang. Authorship attribution with topic drift model. In *The Thirty-First AAAI Conference on Artificial Intelligence*, pages 1–4. AAAI, 2017.

[Zhu *et al.*, 2012] Jia Zhu, Qing Xie, and Eun Jung Chin. A hybrid time-series link prediction framework for large social network. *Database and Expert Systems Applications*, pages 345–359, 2012.

[Zhu *et al.*, 2016] Jia Zhu, Qing Xie, Shoou-i Yu, and Wai Hung Wong. Exploiting link structure for web page genre identification. *Data Mining and Knowledge Discovery*, 30(3):550–575, 2016.