Accepted Manuscript

Large-scale structural and textual similarity-based mining of knowledge graph to predict drug-drug interactions

Ibrahim Abdelaziz, Achille Fokoue, Oktie Hassanzadeh, Ping Zhang, Mohammad Sadoghi

PII: S1570-8268(17)30029-X
DOI: http://dx.doi.org/10.1016/j.websem.2017.06.002
Reference: WEBSEM 439

To appear in: *Web Semantics: Science, Services and Agents on the World Wide Web*

Received date : 24 October 2016
Revised date : 8 March 2017
Accepted date : 2 June 2017

Please cite this article as: I. Abdelaziz, A. Fokoue, O. Hassanzadeh, P. Zhang, M. Sadoghi, Large-scale structural and textual similarity-based mining of knowledge graph to predict drug-drug interactions, *Web Semantics: Science, Services and Agents on the World Wide Web* (2017), http://dx.doi.org/10.1016/j.websem.2017.06.002

Large-Scale Structural and Textual Similarity-Based Mining of Knowledge Graph to Predict Drug-Drug Interactions

Ibrahim Abdelaziz[a,*], Achille Fokoue[b], Oktie Hassanzadeh[b], Ping Zhang[b], Mohammad Sadoghi[c]

[a] *King Abdullah University of Science & Technology, KSA*
[b] *IBM T.J. Watson Research Center, Yorktown Heights, NY, USA*
[c] *Department of Computer Science, Purdue University, West Lafayette, IN, USA*

Abstract

Drug-Drug Interactions (DDIs) are a major cause of preventable Adverse Drug Reactions (ADRs), causing a significant burden on the patients' health and the healthcare system. It is widely known that clinical studies cannot sufficiently and accurately identify DDIs for new drugs before they are made available on the market. In addition, existing public and proprietary sources of DDI information are known to be incomplete and/or inaccurate and so not reliable. As a result, there is an emerging body of research on in-silico prediction of drug-drug interactions. In this paper, we present Tiresias, a large-scale similarity-based framework that predicts DDIs through link prediction.Tiresias takes in various sources of drug-related data and knowledge as inputs, and provides DDI predictions as outputs. The process starts with semantic integration of the input data that results in a knowledge graph describing drug attributes and relationships with various related entities such as enzymes, chemical structures, and pathways. The knowledge graph is then used to compute several similarity measures between all the drugs in a scalable and distributed framework. In particular, Tiresias utilizes two classes of features in a knowledge graph: local and global features. Local features are derived from the information directly associated to each drug (i.e., one hop away) while global features are learnt by minimizing a global loss function that considers the complete structure of the knowledge graph. The resulting similarity metrics are used to build features for a large-scale logistic regression model to predict potential DDIs. We highlight the novelty of our proposed Tiresias and perform thorough evaluation of the quality of the predictions. The results show the effectiveness of Tiresias in both predicting new interactions among existing drugs as well as newly developed drugs.

Keywords: Drug Interaction, Similarity-Based, Link Prediction

1. Introduction

Adverse drug reactions (ADRs) is now becoming the 4^{th} leading cause of deaths in United States surpassing complex diseases such as diabetes, pneumonia, and AIDS [1]. Over two million ADRs are being reported in U.S. annually that sadly results in 100,000 loss of life every year. Furthermore, a significant resource of $136 billion is dedicated to treat complications arised due to ADRs. In fact, the cost of care for attempting to reverse ADRs symptoms is higher than the cost of care for both diabetic and cardiovascular combined. More importantly, a detailed analysis of ADR incidents reveals that approximately 3 to 5% of all in-hospital medication errors are due to "preventable" drug-drug interactions (DDIs) [1].

Therefore, a natural question arises as to why so many preventable DDIs continues to plaque patients and the healthcare system as a whole, the answer is twofold. First, despite the advances made in drug development and safety, clinical trails often fail to reveal rare toxicity of certain drugs given the limited size and length of these studies. For instance, an average typical trail for any drug is limited to only 1,500 patients for rather a short period of time. Therefore, they fail to show the actual impacts of the drug once offered to millions of patients for much longer period of time. These concerns are further exacerbated as it is well known that adverse reaction increases exponentially when taking four or more drugs simultaneously [1]. Consequently, the rare toxicity of newly developed drugs cannot be established until after the drug becomes widely available in the market. Second, to make the matter worse, healthcare providers often fail to report ADRs because they have a misconception that all severe adverse reactions are already known when a drug is brought to the market [1].

Recently, there is a growing interest in computationally predicting potential DDIs [2, 3, 4, 5, 6, 7, 8]. These approaches are broadly classified as either similarity (e.g., [2, 6, 7]) or feature-based (e.g., [3]) DDI predication methods. There are a set of significant challenges and shortcomings that that are mostly overlooked by prior work. We summarize each of these limitations as follows:

Problem 1: Inability to make predictions for newly developed drugs. Prior work either (*i*) are fundamentally unable

*Corresponding author
Email addresses: ibrahim.abdelaziz@kaust.edu.sa (Ibrahim Abdelaziz), achille@us.ibm.com (Achille Fokoue), hassanzadeh@us.ibm.com (Oktie Hassanzadeh), pzhang@us.ibm.com (Ping Zhang), msadoghi@purdue.edu (Mohammad Sadoghi)