Predicting Drug-Drug Interactions Through Similarity-Based Link Prediction Over Web Data

Achille Fokoue, Oktie Hassanzadeh, Mohammad Sadoghi, Ping Zhang
IBM T.J. Watson Research Center
{afokoue,hassanzadeh,msadoghi,pzhang}@us.ibm.com

ABSTRACT

Drug-Drug Interactions (DDIs) are a major cause of preventable adverse drug reactions and a huge burden on public health and the healthcare system. On the other hand, there is a large amount of drug-related (open) data published on the Web, describing various properties of drugs and their relationships to other drugs, genes, diseases, and related concepts and entities. In this demonstration, we describe an end-to-end system we have designed to take in various Web data sources as input and provide as output a prediction of DDIs along with an explanation of why two drugs may interact. The system first creates a knowledge graph out of input data sources through large-scale semantic integration, and then performs link prediction among drug entities in the graph through large-scale similarity analysis and machine learning. The link prediction is performed using a logistic regression model over several similarity matrices built using different drug similarity measures. We present both the efficient link prediction framework implemented in Apache Spark, and our APIs and Web interface for predicting DDIs and exploring their potential causes and nature.

Keywords

Drug-Drug Interactions, Link Prediction, Semantic Web, Big Data on the Web

1. INTRODUCTION

Adverse drug reactions (ADRs) are a major cause of serious health complications and a major burden on the healthcare system. Drug-Drug Interactions (DDIs) are among the leading causes of "preventable" ADRs, in part due to the extreme difficulty of identifying potential DDIs early in the drug design process and through clinical studies. On the other hand, there is a wealth of information on the Web related to drugs, with several sources publishing structured and semi-structured data on the Web. A recent study has shown that none of the existing public sources that contain DDI information provide a reasonable coverage of all the known interactions [2], and most sources are either incomplete or too conservative by listing a large number of insignificant DDIs. Furthermore, existing sources rarely provide an evidence or an explanation for a DDI. As a result, a simple integration of all the public sources would be far from usable in real clinical and pharmaceutical settings.

In this paper, we present Tiresias, a system that takes in as input a variety of drug-related data sources from the Web, including a (small) set of known DDIs, and provides as output a list of potential (unknown) DDIs along with an explanation for each DDI. The system first performs a semantic integration of the input data, building a knowledge graph describing drugs and connecting them to various related entities such as enzymes, chemical structures, and pathways. Similar to content-based recommender systems, the prediction an interaction between a candidate pair of drugs is performed by comparing it against known interacting pairs of drugs. A large number of robust similarity measures taking into account the properties of various sources are calculated and then used to build a linear regression learning model, all in a highly scalable way implemented in Apache Spark.

In what follows, we describe the overall architecture of the system along with a brief summary of various novel solutions implemented in the system to perform semantic integration, feature engineering, and large-scale learning suitable for sparse features gathered from various information sources on the Web. We then present a summary of our demonstration plan.

2. SYSTEM OVERVIEW

The overall architecture of our similarity-based DDI prediction approach is illustrated in Figure 1. In what follows, we briefly describe each of the components

2.1 Knowledge Curation

We construct a knowledge graph by ingesting data from variety of sources (including XML, relational, and CSV formats) from the Web. As partially shown in Figure 2, our data comes from variety of sources such as *DrugBank* [11] that offers data about known drugs and diseases, *Comparative Toxicogenomics Database* [7] that provides information about gene interaction, *Uniprot* [1] that provides details about the functions and structure of genes, *BioGRID* database that collects genetic and protein interactions [6], *Unified Medical Language System* that one is the largest repository of biomedical vocabularies including *NCBI* taxonomy, *Gene Ontology (GO)*, the *Medical Subject Headings (MeSH)* [3], and the *National Drug File - Reference Termi-*

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890532.

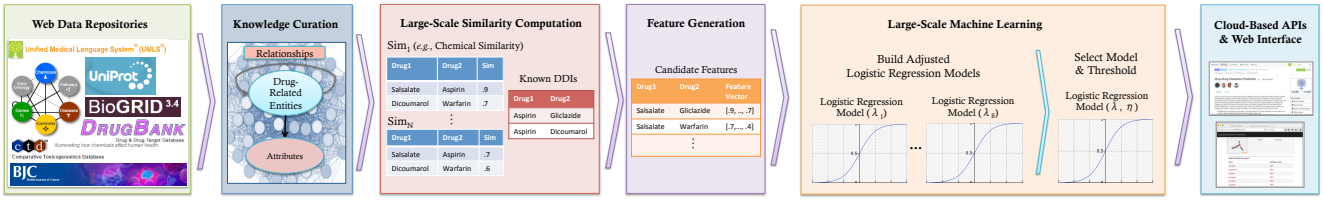

Figure 1: Tiresias **System Architecture.**

nology (NDF-RT) that classifies drug with a multi-category reference models such as cellular or molecular interactions and therapeutic categories [4].

As part of our knowledge graph curation task, we identify which attributes or columns refer to which real world entities (i.e., data instances). Therefore, our constructed knowledge graph possess a clear notion of what the entities are, and what relations exist for each instance in order to capture the data interconnectedness. These may be relations to other entities, or the relations of the attributes of the entity to data values. As an example, in our ingested and curated data, we have a table for *Drug*, and have the columns *Name*, *Targets*, *Symptomatic Treatment*. Our knowledge graph has an identifier for a real world drug *Methotrexate*, and captures its attributes such as *Molecular Structure* or *Mechanism of Actions*, as well as relations to other entities including *Genes* that *Methotrexate* targets (e.g., *DHFR*), and subsequently, *Conditions* that it treats such as *Osteosarcoma (bone cancer)* that are reachable through its target genes, as demonstrated in Figure 2. Constructing a rich knowledge graph is a necessary step before building our predication model as discussed next.

2.2 Similarity Computation

In this phase, data originating from multiple sources that are integrated in our knowledge graph are used to create various drug similarity measures (represented as blue tables in Figure 1) and a known DDIs table. Similarity measures are not necessarily complete in the sense that some drug pairs may be missing from the similarity tables. The known DDIs table, denoted $KDDI$, contains the set of 12,104 drug pairs already known to interact in DrugBank. In the 10-fold cross validation of our approach, $KDDI$ is randomly split into 3 disjoint subsets: $KDDI_{train}$, $KDDI_{val}$, and $KDDI_{test}$ representing the set of positive examples respectively used in the training, validation and testing (or prediction) phases. Contrary to most prior work, which partition $KDDI$ on the DDI associations instead of on drugs, our partitioning simulates the scenario of the introduction of newly developed drugs for which no interacting drugs are known. In particular, each pair (d_1, d_2) in $KDDI_{test}$ is such that either d_1 or d_2 does not appear in $KDDI_{train}$ or KDD_{val}.

Due to space limitation, we describe here only 4 of the 13 similarity measures used to compare two drugs. The other similarity metrics are presented in detail in [8], including physiological effect based similarity, side effect based similarity, two metabolizing enzyme based similarities, three drug target based similarities, chemical structure similarity, MeSH based similarity. Note that the similarity computation using all the measures over all the possible pairs of drugs is quite expensive, and so we utilize Apache Spark for an efficient parallel similarity computation.

- **Chemical-Protein Interactome (CPI) Profile based Similarity:** The Chemical-Protein Interactome (CPI) profile of a drug d, denoted $cpi(d)$, is a vector indicating how well its chemical structure docks or binds with about 611 human Protein Data Bank (PDB) structures associated with DDIs [12]. The CPI profile based similarity of two drugs d_1 and d_2 is computed as the cosine similarity between the mean-centered versions of vectors $cpi(d_1)$ and $cpi(d_2)$.

- **Mechanism of Action based Similarity:** For a drug d, we collect all its mechanisms of action obtained from NDF-RT. To discount popular terms, Inverse Document Frequency (IDF) is used to assign more weight to relatively rare mechanism of actions: $IDF(t, Drugs) = log\frac{|Drugs|+1}{DF(t,Drugs)+1}$ where $Drugs$ is the set of all drugs, t is a mechanism of action, and $DF(t, Drugs)$ is the number of drugs with the mechanism of action t. The IDF-weighted mechanism of action vector of a drug d is a vector $moa(d)$ whose components are mechanisms of action. The value of a component t of $moa(d)$, denoted $moa(d)[t]$, is zero if t is not a known mechanism of action of d; otherwise, it is $IDF(t, Drugs)$. The mechanism of action based similarity measure of two drugs d_1 and d_2 is the cosine similarity of the vectors $moa(d_1)$ and $moa(d_2)$.

- **Pathways based Similarity:** Information about pathways affected by drugs is obtained from CTD database. The pathways based similarity of two drugs is defined as the cosine similarity between the IDF-weighted pathways vectors of the two drugs, which are computed in a similar way as IDF-weighted mechanism of action vectors.

- **Anatomical Therapeutic Chemical (ATC) Classification System based Similarity:** ATC [13] is a classification of the active ingredients of drugs according to the organs that they affect as well as their chemical, pharmacological and therapeutic characteristics. The classification consists of multiple trees representing different organs or systems affected by drugs, and different therapeutical and chemical properties of drugs. The ATC codes associated with each drug are obtained from DrugBank. For a given drug, we collect all its ATC code from DrugBank to build a ATC code vector (the most specific ATC codes associated with the drug -i.e., leaves of the classification tree- and also

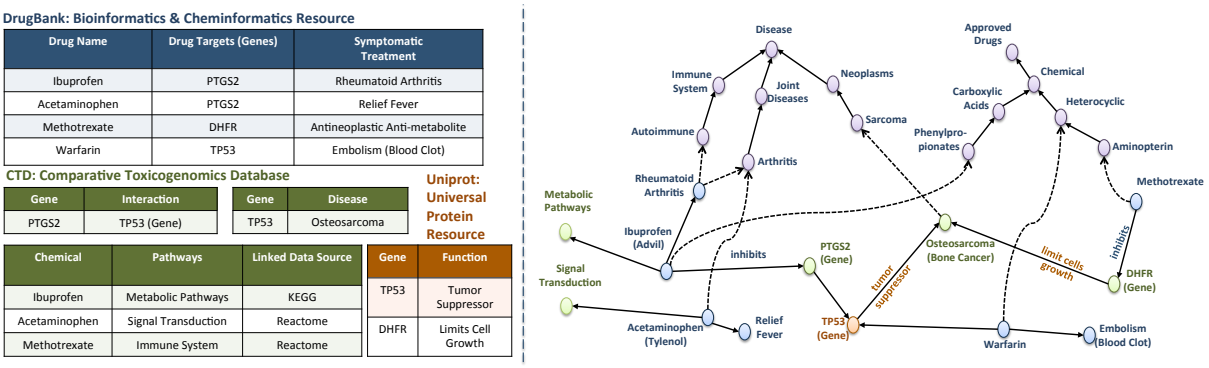

Figure 2: Semantic Curation and Linkage of Data from Variety of Sources on the Web.

all the ancestor codes are included). The ATC similarity of two drugs is defined as the cosine similarity between the IDF-weighted ATC code vectors of the two drugs, which are computed in a similar way as IDF-weighted mechanism of action vectors.

2.3 Feature Generation

Given a pair of drugs (d_1, d_2), we construct its machine learning feature vector derived from the drug similarity measures and the set of DDIs known at training. Like previous similarity-based approaches, for a drug candidate pair (d_1, d_2) and a drug-drug similarity measure $sim_1 \otimes sim_2$, we create a feature that indicates the similarity value of the known pair of interacting drugs most similar to (d_1, d_2). Unlike prior work, we introduce new calibration features to address the issue of the incompleteness of the similarity measures and to provide more information about the distribution of the similarity values between a drug candidate pair and all known interacting drug pairs - not just the maximum value.

2.4 Large-Scale Machine Learning

Our learning framework consists of three phases, all implemented on top of Apache Spark for scalability:

- **Model Validation Phase** As a result of relying on more data sources, using more similarity measures, and introducing new calibration features, we have significantly more features (1014) than prior work (e.g., [9] uses only 49 features). Thus, there is an increased risk of overfitting that we address by performing L_2-model regularization. Since the optimal regularization parameter is not known a-priori, in the model generation phase, we build 8 different logistic regression models using 8 different regularization values. To address issues related to the skewed distribution of DDIs (for an assumed prevalence DDIs lower than 17%), we make some adjustments to logistic regression.

- **Model Validation Phase** The goals of this phase are twofold. First, in this phase, we select the best of the eight models (i.e., the best regularization parameter value) built in the model generation phase by choosing the model producing the best F-score on the validation data. Second, we also select the optimal threshold as the threshold at which the best F-score is obtained on the validation data evaluated on the selected model.

- **Prediction Phase** Let f denote the logistic function selected in the model validation phase and η the confidence threshold selected in the same phase. In the prediction phase, for each candidate drug pair (d_1, d_2), we first get its feature vector v computed in the feature construction phase. $f(v)$ then indicates the probability that the two drugs d_1 and d_2 interact, and the pair (d_1, d_2) is labeled as interacting iff. $f(v) \geq \eta$.

2.5 APIs & Web Interface

We provide the outcome of our similarity-based DDI prediction through a set of APIs. These APIs not only provide the predicted DDIs along with confidence scores and references, but also provide access to the internals of the system so the users can query and analyze the reason a DDI shows up in the output. As a result, the outcome can be integrated within existing Electronic Health Record (EHR) systems along with features that would assist clinicians and healthcare professionals in not only providing up-to-date DDI information, but also analyzing the reason behind a particular DDI and the associated risk. We have also designed a set of Web interfaces over our APIs (e.g., see Figure 3).

3. DEMONSTRATION PLAN

Our plan is to demonstrate three aspects of Tiresias:

- **Knowledge Graph Curation** We allow the users to navigate through our drug knowledge graph (cf. Figure 2) in RDF through the popular LodLive interface [5]. We use custom URIs that mark each predicate with the source(s) of each fact and use that to show the different coverage of various Web sources with respect to different types of properties and relationships. This portion of the demonstration would be of interest to the general Semantic Web and Linked Data audience, pointing out challenges in building healthcare applications using disparate Web sources and the shortcomings of the existing sources published as Linked Data on the Web [10].

- **Similarity Analysis Component** Through a Web interface, users can select two drugs and see the similarity computation results using all the 13 similarity measures we have implemented in Tiresias. We use this interface to show the usefulness of each the similarity measures in capturing different types of similarities

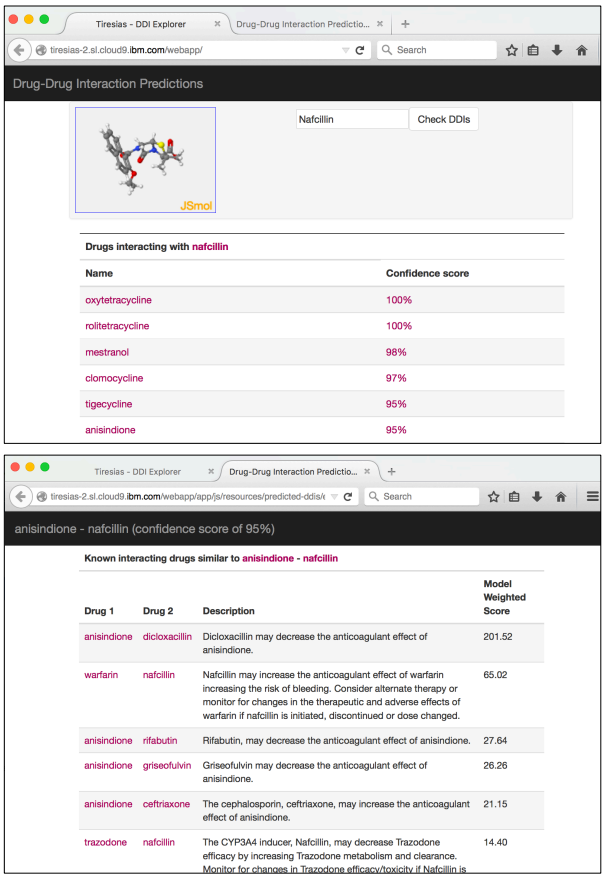

Figure 3: DDI Exploration Web Interface

across drugs that are useful for DDI prediction. We also show how the results could be incomplete for certain DDIs due to the different coverage of the input sources.

- **DDI Predictions & Explanations** The main part of our demonstration will be showcasing the end result of DDI predictions. First, we present a Web interface to look up known and predicted interactions for a given drug, and provide evidence and explanation for each DDI as shown in Figure 3. The explanation is based on the existing evidence for interaction between similar drugs. We then present the results of retrospective analysis which relies on DDI information from an older version of DrugBank to predict interactions that were unknown at the time, verify some new interactions that have been discovered since then, and compare the explanation our system provides with the new evidence supporting the existence of the DDI.

Our demonstration will be using a prototype cloud service that is currently at Alpha stage. In addition to the above, we will share the current technical documentation, API design, business use cases, and our plan for commercial offering of the solution.

4. REFERENCES

[1] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, et al. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl 1):D115–D119, 2004.

[2] S. Ayvaz, J. R. Horn, O. Hassanzadeh, Q. Zhu, J. Stan, N. P. Tatonetti, S. Vilar, M. Brochhausen, M. Samwald, M. Rastegar-Mojarad, M. Dumontier, and R. D. Boyce. Toward a complete dataset of drug-drug interaction information from publicly available sources. *Journal of Biomedical Informatics*, 55:206–217, 2015.

[3] O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[4] S. H. Brown, P. L. Elkin, S. Rosenbloom, C. Husser, B. Bauer, M. Lincoln, J. Carter, M. Erlbaum, and M. Tuttle. VA National Drug File Reference Terminology: a cross-institutional content coverage study. *Medinfo*, 11(Pt 1):477–81, 2004.

[5] D. V. Camarda, S. Mazzini, and A. Antonuccio. LodLive: exploring the web of data. In *International Conference on Semantic Systems (I-SEMANTICS '12), Graz, Austria, September 5-7, 2012*, pages 197–200, 2012.

[6] A. Chatr-aryamontri, B.-J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D. Chen, C. Stark, A. Breitkreutz, N. Kolas, L. O'Donnell, et al. The BioGRID interaction database: 2015 update. *Nucleic acids research*, page gku1204, 2014.

[7] A. P. Davis, C. G. Murphy, C. A. Saraceni-Richards, M. C. Rosenstein, T. C. Wiegers, and C. J. Mattingly. Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic acids research*, 37(suppl 1):D786–D792, 2009.

[8] A. Fokoue, M. Sadoghi, O. Hassanzadeh, and P. Zhang. Predicting drug-drug interactions through large-scale similarity-based link prediction. `http://ibm.biz/adrtechreport`.

[9] A. Gottlieb, G. Y. Stein, Y. Oron, E. Ruppin, and R. Sharan. Indi: a computational framework for inferring drug interactions and their associated recommendations. *Molecular systems biology*, 8(1):592, 2012.

[10] A. Jentzsch, J. Zhao, O. Hassanzadeh, K. Cheung, M. Samwald, and B. Andersson. Linking open drug data. In *5th International Conference on Semantic Systems, Graz, Austria, September 2-4, 2009. Proceedings*, 2009.

[11] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, et al. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(suppl 1):D1035–D1041, 2011.

[12] H. Luo, P. Zhang, H. Huang, J. Huang, E. Kao, L. Shi, L. He, and L. Yang. Ddi-cpi, a server that predicts drug–drug interactions through implementing the chemical–protein interactome. *Nucleic acids research*, page gku433, 2014.

[13] A. Skrbo, B. Begović, and S. Skrbo. [classification of drugs using the atc system (anatomic, therapeutic, chemical classification) and the latest changes]. *Medicinski arhiv*, 58(1 Suppl 2):138–141, 2003.