Mining Patterns from Clinical Trial Annotated Datasets by Exploiting the NCI Thesaurus

Joseph Benik[1], Guillermo Palma[2], Louiqa Raschid[1], Andreas Thor[1], and Maria-Esther Vidal[2]

[1] University of Maryland, College Park, USA
josephbenik@gmail.com, {louiqa,thor}@umiacs.umd.edu
[2] Universidad Simón Bolívar, Venezuela
{gpalma, mvidal}@ldc.usb.ve

Abstract. Annotations of clinical trials with controlled vocabularies of drugs and diseases, encode scientific knowledge that can be mined to discover relationships between scientific concepts. We present PAnG (Patterns in Annotation Graphs), a tool that relies on dense subgraphs, graph summarization and taxonomic distance metrics, computed using the NCI Thesaurus, to identify patterns.

1 Introduction

Linked Open Data has made available a diversity of collections and can facilitate scientists to mine semantically annotated datasets. These annotations represent scientific knowledge, for example, genes, proteins, drugs and diseases are annotated with controlled vocabulary terms (CV terms) from ontologies. Annotations describe properties of these concepts, and they are useful as a basis for focused literature review, and further, to plan a wet-lab experiment or a clinical trial. Annotation graphs as well as the ontologies are rich and complex, for example, the NCI Thesaurus version 12.05d has 93,788 terms. Thus, the challenge is to explore the potentially large number of annotations and to discover patterns. Automatic techniques and tools are therefore needed to support the scientist. These tools could range from making simple but valuable link predictions, e.g., predicting new gene functional annotation, to discovering complex patterns of annotation across multiple disease conditions and drug interventions.

We present PAnG (Patterns in Annotation Graphs), a tool that allows scientists to identify patterns in annotated graph datasets. PAnG is based on a complementary methodology of graph summarization (GS) and dense subgraphs (DSG) [3, 4]. DSG shows particular benefit in creating a promising subgraph, when the input graph is large and includes a diversity of ontology terms, or when the graph has sparse annotations. PAnG uses a taxonomic distance metric, d_{tax} [2] to compute distances between terms, e.g., in the NCI Thesaurus. Patterns are represented as graph summaries that consist of node partitions (super-nodes). Patterns can include super-edges between super-nodes as well as edges between individual nodes. Patterns provide a better visualization and understanding of the overall structure of the underlying graph.

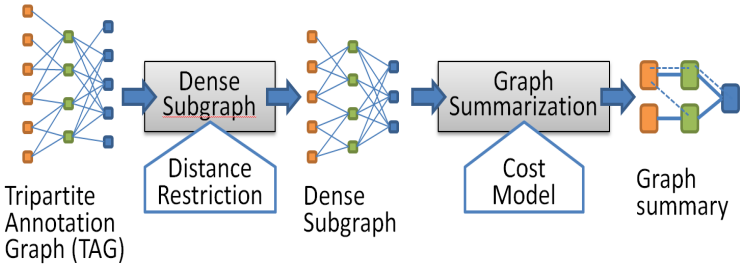

Fig. 1. The PAnG System Workflow.

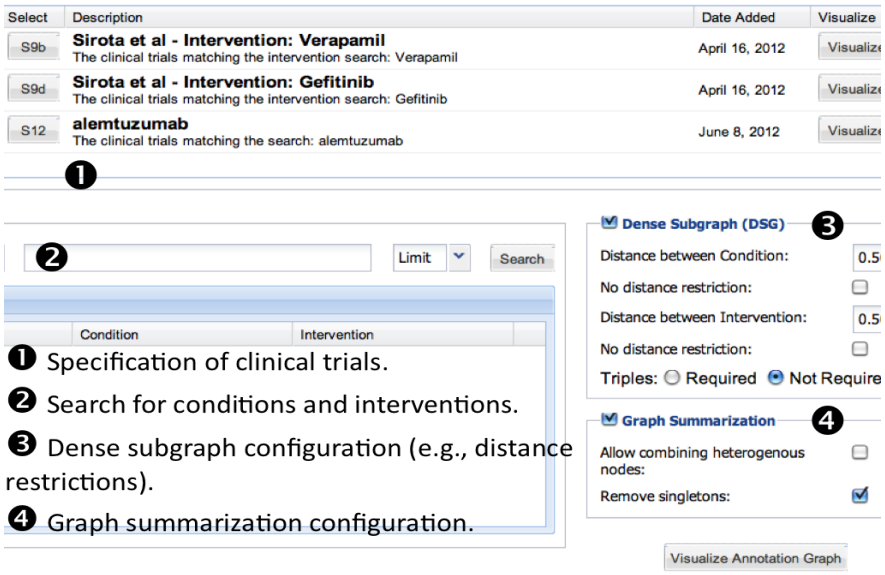

Fig. 2. PAnG's GUI for the LinkedCT dataset.

Further, the pattern captures semantic knowledge not only about individual nodes and their connections, but also about groups of related nodes. LinkedCT (LinkedCT.org) is a Linked Open Data dataset from the clinical trial site (ClinicalTrial.gov). Conditions represent diseases and are typically described using the NCI Thesaurus. Interventions include a (unique) name, a description and a type, e.g., a drug, device, procedure, etc. PAnG for LinkedCT.org is available at `http://pang.umiacs.umd.edu/iswc2012demo`.

2 The PAnG System

Figure 1 illustrates the overall workflow of PAnG. The input is a tripartite annotated graph G, and the output is a graph summary. Our workflow consists of two steps. The first step is optional and deals with the identification of dense subgraphs, i.e., highly connected subgraphs of G that are (almost) cliques. The

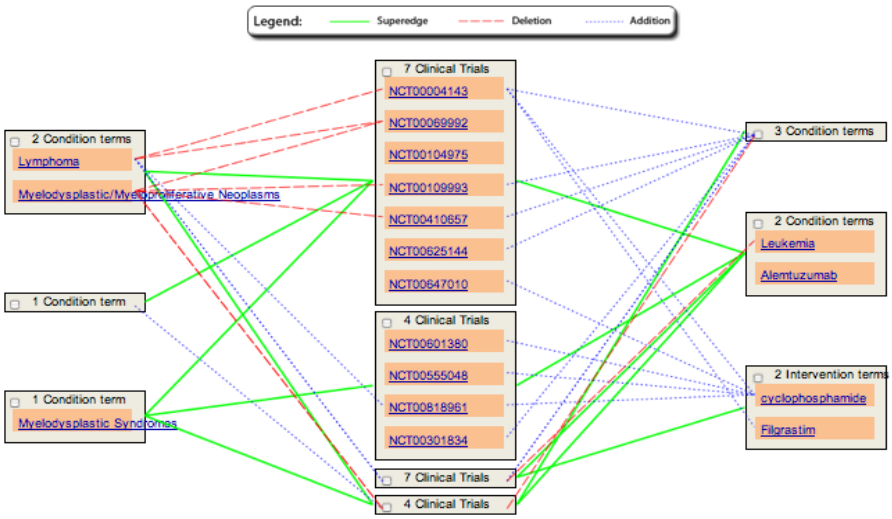

Fig. 3. Graph Summary for Alemtuzumab. PAnG configuration: S12 (alemtuzumab), DSG+GS, Triples Not Required, Distance between Condition: 0.5, Distance between Intervention: 0.5, Allowing combining heterogenous nodes, Remove singletons.

goal is to identify interesting regions of the graph by extracting a relevant subgraph. Graph summarization transforms the graph into an equivalent compact graph representation. Graph summaries are made up of the following elements: (1) supernodes; (2) superedges; (3) deletion and addition edges (corrections). Figure 2 shows the PAnG interface for the LinkedCT annotation graph dataset. A scientist can search on conditions or interventions and create a subset of clinical trials. She can then choose the DSG option, with a threshold for d_{tax} for conditions or interventions or both. She can also skip the DSG option or choose a DSG without a distance restriction. She can also require that the DSG option favor the selection of triplets across the tri-partite graph, or favor the selection of independent doublets across the two bi-partite graphs. Figure 3 presents a possible summary graph. There are 10 supernodes; five supernodes cluster conditions, four supernodes cluster clinical trials and one supernode clusters the two interventions cyclophosphamide, and Filgrastim. A superedge is a solid edge and occurs between two supernodes. It represents that all nodes in both supernodes are fully connected, for example, the superedge between the supernode with the condition Myelodysplastic Syndromes and the supernode of 4 clinical trials. The summary reflects the basic pattern (structure) of the graph and is accompanied by a list of corrections, i.e., deletions and additions, that express differences between the graph and its simplified pattern. For example, a deletion edge to a condition that occurs within a supernode indicates that the specific condition *was not* studied in a particular clinical trial, whereas the conditions within the supernode, without deletion edges, were studied. In Figure 3, the condition Lymphoma (top left-hand supernode) was not studied in the clinical trial NCT00004143 $\cdots$ (top middle supernode).

3 Demonstration of Use Cases

As of September 2011, LinkedCT contains 106,308 trials, 2.7 million entities and over 25 million RDF triples. We demonstrate two use cases:
Single Drug: We commence with a single drug and create a dataset of all clinical trials associated with that drug, and all other conditions and interventions associated with these trials. We use the taxonomic distance metric, d_{tax} [1], and the NCI Thesaurus version 12.05d to compute pairwise (condition-condition) or (intervention-intervention) distance. Because, this is a poly-hierarchy, if there are alternate paths, the shortest path is chosen. Three drugs Alemtuzumab, Getfinib, and Verapamil are used to illustrate the effect of different configurations. For example, Getfinib treats certain types of cancers, e.g., breast or lung cancer. With no DSG, PAnG cannot discern any patterns. If DSG+GS are chosen, with no distance restriction and triples required, PAnG produces a very simple summary of a supernode of clinical trials for Breast Cancer, another supernode for Lung Cancer, with only Getfinib as the intervention, and one clinical trial covering both cancers. However, with DSG+GS, no distance restriction, no triples required, a different graph summary is generated. For example, with a distances threshold of 0.3 for both conditions and interventions, Esophageal Cancer is related to a supernode of Radiation Therapy, conventional surgery and neoadjuvant therapy; this suggests that this disease is related to these three procedures, in addition to treatments with Getfinib.
Drug Family: The NCI Thesaurus is used to explore drug families and their properties. Starting with Alemtuzumab as an exemplar, we retrieve the intersection of Monoclonal antibodies and Antineoplastic agents. This creates a dataset of 12 drugs: Alemtuzumab, Bevacizumab, Brentuximab vedotin, Cetuximab, Catumaxomab, Edrecolomab, Gemtuzumab, Ipilimumab, Ofatumumab, Panitumumab, Rituximab, and Trastuzumab. We use the pairwise annotation similarity based on the set of interventions associated with each drug to select interesting pairs of drugs[2] and then further analyze using PAnG.

References

1. J. Benik, C. Chang, L. Raschid, M.-E. Vidal, G. Palma, and A. Thor. Finding cross genome patterns in annotation graphs. In *DILS*, pages 21–36, 2012.
2. G. Palma, E. Haag, L. Raschid, A. Thor, and M.-E. Vidal. An Evaluation of Metrics to Compute Concept Similarity Based on Evidence from Ontologies. *Technical Report, University of Maryland UMIACS*, 2012.
3. B. Saha, A. Hoch, S. Khuller, L. Raschid, and X.-N. Zhang. Dense subgraphs with restrictions and applications to gene annotation graphs. In *Conference on Research on Computational Molecular Biology (RECOMB)*, 2010.
4. A. Thor, P. Anderson, L. Raschid, S. Navlakha, B. Saha, S. Khuller, and X.-N. Zhang. Link prediction for annotation graphs using graph summarization. In *Proc. of International Semantic Web Conference (ISWC)*, 2011.