PAnG - Finding Patterns in Annotation Graphs

Philip Anderson
University of Maryland
phand@umd.edu

Andreas Thor
University of Maryland
thor@umiacs.umd.edu

Joseph Benik
University of Maryland
josephbenik@gmail.com

Louiqa Raschid
University of Maryland
louiqa@umiacs.umd.edu

Maria Esther Vidal
Universidad Simon Bolivar
mvidal@ldc.usb.ve

ABSTRACT

Annotation graph datasets are a natural representation of scientific knowledge. They are common in the life sciences and health sciences, where concepts such as genes, proteins or clinical trials are annotated with controlled vocabulary terms from ontologies. We present a tool, PAnG (Patterns in Annotation Graphs), that is based on a complementary methodology of graph summarization and dense subgraphs. The elements of a graph summary correspond to a pattern and its visualization can provide an explanation of the underlying knowledge. Scientists can use PAnG to develop hypotheses and for exploration.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Graphs and Networks;
H.2.8 [**Database Applications**]: Data Mining

Keywords

Link Prediction, Graph Summarization, Dense Subgraphs

1. INTRODUCTION

The Linking Open Data (LOD) initiative has been successful in providing access to a diversity of data collections. LOD can facilitate transformative applications that allow scientists to share and mine richly annotated and/or extensively hyperlinked collections. Annotation graph datasets are a natural representation of scientific knowledge; for example, genes, proteins and clinical trials are annotated with controlled vocabulary terms (CV terms) from ontologies. Figure 1 (left) illustrates such a tripartite annotation graph for the gene CRY2 using terms from the Plant Ontology (PO) and Gene Ontology (GO). Ontologies capture multiple relationship types between CV terms. Figure 1 (right) shows a fragment of the Gene Ontology and relationship types part-of, is-a, and regulates.

Sensemaking from annotation graphs is useful to scientists since it can help in planning a wet-lab experiment or serve as a basis for focused literature review. Since annotation graphs can be huge – the Gene Ontology itself contains more than 35,000 terms as of October 2011 – the challenge for the scientist is to explore the potentially large number of annotations and to discover useful patterns. Automatic techniques and tools are therefore needed to support the scientist. Sensemaking 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 that can lead to important scientific advances.

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) [5, 7, 8]; these methodologies were developed by the researchers and their collaborators. DSG uses the ontology structure, in particular the distance between CV terms, to filter the graph. 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. Patterns are represented as graph summaries that consist of node partitions (supernodes). Patterns can include superedges between supernodes as well as edges between individual nodes. Patterns provide a better visualization and understanding of the overall structure of the underlying graph. Further, the pattern captures semantic knowledge not only about individual nodes and their connections but also about groups of related nodes. PAnG offers the following key features:

- Exploration and identification of an annotation graph dataset (including relevant ontology fragments).
- Identification of dense subgraphs with constraints on the distances between pairs of ontology terms.
- Visualization of a pattern represented by a graph summary, i.e., a hypergraph with supernodes, superedges, and corrections (to be discussed later).

2. OVERVIEW OF PANG

Figure 2 illustrates the overall workflow of PAnG. The input is typically 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 goal is to identify interesting regions of the graph by extracting a relevant subgraph. Examples of the parameters that can be used to select subgraphs are presented in the next section.

SIGMOD '12, May 20–24, 2012, Scottsdale, Arizona, USA.

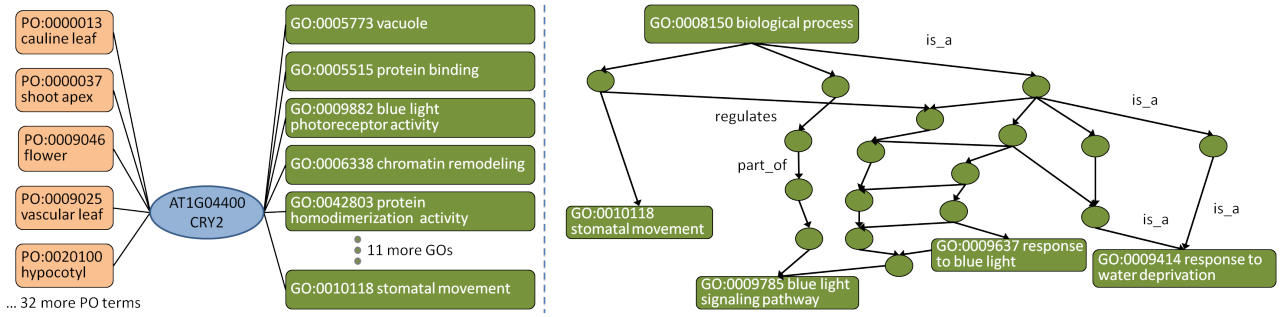

Figure 1: Annotation graph for gene CRY2 using terms from Plant Ontology (PO) and Gene Ontology (GO) (left); fragment of the Gene Ontology incl. different relationships between terms (right).

Next, 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 3 shows a screenshot of a possible summary of an annotation graph. For example, there are 3 gene supernodes (middle layer) and the top gene supernode includes the three 3 genes CHX23, CHX10, and CHX28. 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 middle PO supernode (with 2 PO terms carpel and sepal) and the middle gene supernode (with 4 genes) indicates that all 4 genes are each annotated with both PO terms.

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 reflects that a gene *does not have* a particular annotation that is shared by other genes within the supernode. In Figure 3, the gene CHX28 (top supernode) is not annotated with the GO term sodium $\cdots$ (right layer).

A graph summary has many benefits. First, the summary provides insight into the overall structure of the underlying graph and is good for visualization. Next, it captures semantic knowledge not only about individual nodes and their connections but also about groups of related nodes. Third, the corrections, in particular deletions, are intuitive indicators for future edge prediction. Tripartite graphs are chosen as a template for several reasons. Bipartite graphs may not convey complex relationships, whereas the relationships captured in graphs with N layers, where N>3, may not be well suited to a graph summary.

Next, we describe the datasets that are currently used by PAnG, the user interface, and implementation details. Details about dense subgraphs and graph summarization are in Sections 3 and 4, respectively, followed by a description of demonstration use case scenarios in Section 5.

2.1 Datasets

The PAnG tool can be applied to any annotation graph dataset. A dataset can be populated in different ways, e.g., through bulk import of XML tagged files or by executing SPARQL queries [2]. For all datasets, PAnG keeps the references to the original data sources, and can therefore link all nodes of the graphs to the corresponding (external) web pages. We discuss two sample datasets, one representing a model plant organism, and the second representing the results of clinical trials in the health sciences.

The first dataset is TAIR[1], the primary source of annotation data for the Arabidopsis thaliana model organism. TAIR is a highly curated portal representing the collective knowledge of the community of Arabidopsis researchers. Each gene in TAIR is annotated with controlled vocabulary terms from the Gene Ontology and Plant Ontology.

For example, the resulting tripartite annotation graph for the gene CRY2 is illustrated in Figure 1 (left). PO terms are on the left and GO terms are on the right of CRY2. As of September 2011, there were 17 GO and 37 PO annotations for CRY2. Figure 1 illustrates partial annotations only due to space constraints. On the right of Figure 1 is a fragment of the relevant GO ontology.

The second dataset is the NIH clinical trial data and the corresponding hyperlinked collection LinkedCT[2] [3]. A clinical trial (CT) is associated with terms representing conditions and interventions. Conditions represent the disease or condition being studied and typically are described using MeSH terms. Interventions include a (unique) name, description and type, e.g., a drug, device, procedure, supplement, etc. Additionally, each CT may be hyperlinked to CV terms from Drugbank, Dailymed, etc. These links may provide information about equivalent drugs or therapies, additional properties or features of a drug, associated genes, GO terms, etc.

2.2 User interface and Implementation

Figure 5 shows an annotated screenshot of PAnG's user interface for the TAIR dataset. The user can specify genes of interest (CT's for the clinical trial data set) in two different ways. She can use a list of predefined set ① or she can select genes based on search results ②. For the latter, a full-text index is built for all nodes (genes, POs, and GOs) and a (gene, PO, GO) triplet is retrieved if any of the nodes matches the search condition.

After selecting a gene subset, the user can vary parameter settings for both the dense subgraph and graph summarization. For dense subgraphs, she may choose distance restrictions for node pairs in the PO and GO hierarchies ③. She can also restrict the hierarchies to certain types of relationships, e.g., is-a or part-of. For graph summarization, she may allow to merge heterogeneous nodes in the same supernode, e.g., nodes from PO and GO ④.

[1]http:// www.arabidopsis.org/
[2]http://clinicaltrials.gov and http://linkedct.org

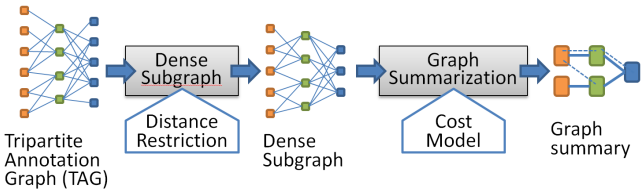

Figure 2: The TAG can pass an optional filter step to identify dense subgraphs. The PAnG tool employs graph summarization to identify patterns.

Note that both steps, dense subgraph and graph summarization, are optional. PAnG can therefore visualize a graph, its summary, a specific dense subgraph, and its summary.

PAnG is made available as a web application [1]. It is written in Java using the Google Web Toolkit. The algorithms for both graph summarization and dense subgraphs are optimized re-implementations of [5] and [7], respectively.

3. DENSE SUBGRAPHS

Given an initial tripartite graph, a challenge is to find interesting regions of the graph, i.e., candidate subgraphs, that can lead to valuable patterns. We commence with the premise that an area of the graph that is rich or dense with annotation is an interesting region to identify candidate subgraphs. For example, for a set of genes, if each is annotated with a set of GO terms and/or a set of PO terms, then the set of genes and GO terms, or the set of genes and PO terms, form a clique. We thus exploit cliques, or dense subgraphs (DSG) representing cliques with missing edges. Density is a measure of the connectedness of a subgraph. It is the ratio of the number of induced edges to the number of vertices in the subgraph.

An annotation graph is a tripartite graph $G = ((A, B, C), (X, Y))$. PAnG employs our approach in [7] and thus first transforms the tripartite graph G in a weighted bipartite graph $G' = (A, C, E)$ where each edge $e = (a, c) \in E$ is labeled with the number of nodes $b \in B$ that have links to both a and c. We then compute a densest bipartite subgraph G_2 by choosing subsets of A and C to maximize the density of the subgraph. Finally, we build the dense tripartite graph G_3 out of the G_2 by adding all intermediate nodes $b \in B$ that are connected to at least one node of G_2.

In the annotated biological web (see Figure 1) nodes from PO and GO are hierarchically arranged to reflect their relationships (e.g., is-a or part-of). The PAnG tool therefore allows users to include restrictions on the ontology terms in the DSG. A distance restriction specifies the maximal distance between pairs of nodes in set A (C). To this end, PAnG employs a distance metric d_A (d_C) and computes the densest subgraph G_3 that ensures that for all node pairs of A (C) are within a given distance τ_A (τ_C). Furthermore, the user can filter the ontology by the relationship type, i.e., only node pairs that are in a specific relationship are considered for distance computation.

4. GRAPH SUMMARIZATION

PAnG generates graph summaries for representing patterns. A summary of a tripartite annotation graph is also a graph. While there are many methods to summarize graphs, we focus on the graph summarization (GS) approach of [5].

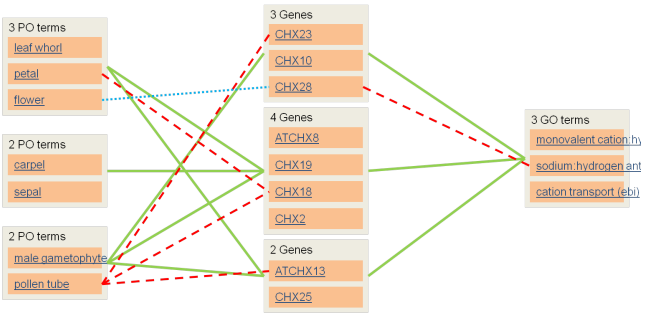

Figure 3: Screenshot of a PAnG graph summary. Superedges are represented by green solid lines; corrections by red dashed (deletion) and blue dotted (addition) lines, resp.

Their graph summary is an aggregate graph comprised of a signature and corrections. It is the first application of minimum description length (MDL) principles to graph summarization and has the added benefit of providing intuitive coarse-level summaries that are well suited for visualization and link prediction.

A graph summary (GS) of a graph $G = ((A, B, C), (X, Y))$ consists of a graph **signature** $\Sigma(G)$ and a set of **corrections** $\Delta(G)$. The graph signature is defined as follows: $\Sigma(G) = ((S_{AC}, S_B), S_{XY})$. The sets S_{AC} and S_B are a disjoint partitioning of $A \cup C$ and B, respectively, that cover all elements of these sets. Each element of S_{AC} or S_B is a **supernode** and consists of one or more nodes of the original graph. An element of S_{XY} is a **superedge** and it represents edges between supernodes, i.e., $S_{XY} \subseteq S_{AC} \times S_B$. The **corrections** are the sets of edge additions and deletions $\Delta(G) = (S_{add}, S_{del})$. All edge additions are edges of the original graph G, i.e., $S_{add} \subseteq X \cup Y$. Deletions are edges between nodes of G that do not have an edge in the original graph, i.e., $S_{Del} \subseteq ((A \cup C) \times B) - (X \cup Y)$. Figure 3 depicts an example summary that we have already discussed in Section 2.

Graph summarization uses a two-part minimum description length encoding and a greedy agglomerative clustering heuristic. The possible summaries of a graph will depend on the cost model used for an MDL encoding. In general, the cost model assigns weights to the number of superedges, deletions, and additions, respectively. Graph summarization looks for a graph summary with a minimal cost. Currently PAnG employs a simple cost model that gives equal weight to supernodes, superedges and corrections.

5. DEMONSTRATION DESCRIPTION

During the tool demonstration we will illustrate how scientists could use PAnG to (1) analyze a known dataset of interest to develop hypotheses, and (2) explore a previously unknown dataset.

5.1 Scenario 1: Hypothesis generation

A scientist typically develops expertise with respect to a set of genes of interest (GOI) within a biological context, e.g., flowering time genes or photomorphogenesis genes. She would typically specify a set of familiar genes to PAnG so that she could identify novel hypotheses, building upon her expertise and the patterns in the graph summaries.

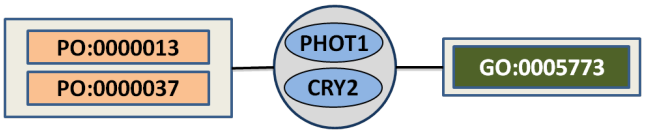

Figure 4: Graph summary for CRY2 and PHOT1.

In our demo, starting with a set of 10 photomorphogenesis genes, we generate the graph summary shown in Figure 4. One supernode contains the two PO terms PO:13 (cauline leaf) and PO:37 (shoot apex), the second includes the two genes CRY2 and PHOT1, and the third has the GO term GO:5773 (vacuole). The three supernodes are connected by two superedges. Subsequently, our scientist may explore the literature to understand the evidence supporting this pattern. As discussed in [4, 6], PHOT1 and CRY2 belong to two different groups of blue light receptors, namely phototropins (PHOT1) and cryptochromes (CRY2). To date there has been *no evidence reported in the literature* to confirm any interactions between these two groups in the vacuole; thus, this is a potential discovery.

Additional annotations included in the dense subgraph and graph summary (not shown) will help the scientist develop a hypothesis. For example, the vacuole is the storage site for ions. When the concentration of ions changes in a guard cell, there will be stomatal movement. PAnG reveals that both genes are annotated with GO terms related to stomatal movement. Further, CRY2 is annotated with the term response to water deprivation; stomata are typically closed during water deprivation. Next, PAnG could highlight that PHOT1 is *not annotated* with this term; for example, it may appear as a deletion in the corresponding graph summary. The scientist would utilize these patterns to develop hypotheses and an experiment to determine whether response to water deprivation should be predicted as a new functional annotation for PHOT1. If true, then PAnG would have identified both a new gene function for PHOT1, and would have isolated the evidence to support the discovery of an interaction between the two genes in the vacuole.

5.2 Scenario 2: Exploration

Over 100,000 clinical trials (CTs) have been made available from LinkedCT circa August 2011. 60,000 CTs are associated with both intervention and condition term(s) and GO and MeSH annotations. A scientist interested in specific drug interventions across a set of disease conditions may have to investigate multiple CTs, their annotations and hyperlinks. For example, 1,109 of the CTs cover breast, colorectal, ovarian and lung cancer. Examples of drug interventions for these cancers include cyclophosphamide, gemcitabine hydrochloride and fludarabine phosphate; cyclophosphamide itself has been studied in over 490 diseases and in 984 CTs. The challenge for the scientist is to explore the potentially large number of annotations and to discover useful patterns.

We illustrate three examples of exploration. The drug intervention everolimus is associated with 260 CTs. Our scientist selects a subset of 100 CTs; an initial graph summary reveals one supernode with 34 CTs. Next, she experiments with several settings of the distance metric for dense subgraphs. She observes that up to 50% of the 34 CTs are conserved in this supernode, over all parameter settings, illustrating that this is a stable (supernode) pattern. An even more significant observation is revealed in her second example with the intervention tracolimus. A supernode of 8 CTs is identified in an initial graph summary. This identical supernode will be conserved over all experiment settings! This indicates that the 8 CTs are closely related to each other. In her final example, she considers the drug intervention paclitaxel carboplatin gefitinib. She will apply DSG+GS (over several experiment conditions) and repeatedly isolate one supernode with 5 CTs, another supernode with the condition lung cancer and a superedge between the 2 supernodes. This suggests that the drug paclitaxel carboplatin gefitinib, the group of 5 CTs and the disease lung cancer share a stable relationship.

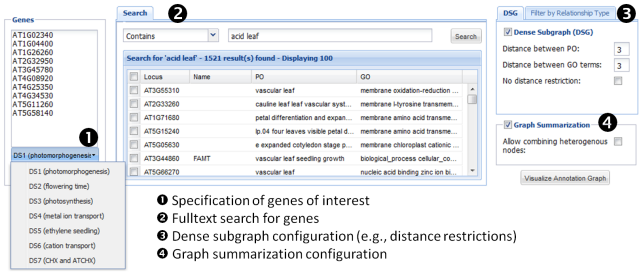

Figure 5: PAnG's GUI for the TAIR dataset.

6. REFERENCES

[1] PAnG. http://pang.umiacs.umd.edu.

[2] M. Acosta, M.-E. Vidal, T. Lampo, J. Castillo, and E. Ruckhaus. Anapsid: An adaptive query processing engine for sparql endpoints. In *Proc. of International Semantic Web Conference (ISWC)*, 2011.

[3] O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang. Linkedct: A linked data space for clinical trials. In *Proc. WWW 2009 Workshop on Linked Data on the Web (LDOW2009)*, 2009.

[4] B. Kang, N. Grancher, V. Koyffmann, D. Lardemer, S. Burney, and M. Ahmad. Multiple interactions between cryptochrome and phototropin blue-light signalling pathways in arabidopsis thaliana. *Planta*, 227(5):1091–1099, 2008.

[5] S. Navlakha, R. Rastogi, and N. Shrivastava. Graph summarization with bounded error. In *Proc. of Conference on Management of Data (SIGMOD)*, 2008.

[6] M. Ohgishi, K. SajiÂ£, K. Okada, and T. Sakai. Functional analysis of each blue light receptor, cry1, cry2, phot1, and phot2, by using combinatorial multiple mutants in arabidopsis. *Proc. of the National Academy of Sciences*, 1010(8):2223–2228, 2004.

[7] 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.

[8] 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.