ABSTRACT
Researchers in academia and industry face a deluge of data in our digital world. In this paper, we investigate a novel problem, query-centric scientific topic evolution. Using heterogeneous graph mining techniques we construct a topic evolution tree (TET) from massive collections of scientific publications, enabling students and researchers to explore the foundation of research topics outside their specialization. Prior research has focused mainly on citation relationships; in this study we employed multiple types of relationships, including authorship, citation, publishing venue, and the contributions authors, papers, and venues have made to a specific topic. We examine multiple restricted meta-paths in constructing a TET covering topics from the MeSH vocabulary for biomedical research.

Keywords
Topic Evolution, MeSH, Heterogeneous Graph.

INTRODUCTION
Scholarly resources are often indexed based on keywords that identify important topics in a domain, and for researchers entering a field or searching for data outside their specialization, such indexes can provide valuable guidance. However, while keywords can identify topics that are important within a domain, they do not provide insight into how topics have evolved or identify which resources or researchers contributed the most to that evolutionary process. The goal of this project is to assist students and researchers by showing the evolution of topics of interest to them; identifying both how topics are related and their evolution. The directed relationship between key topics in a domain can be extracted from a heterogeneous graph composed of the articles, researchers, and scientific venues and their ranked contribution to future topics.

Prior research on topic evolution has leveraged Latent Dirichlet Allocation (Blei et al., 2003) to extract topics from a corpus and predict future topics, while research on heterogeneous graphs has focused on classification, clustering, recommendation, and citation recommendation (Sun & Han, 2012, Yu et al., 2014, Liu et al. 2014). This research extends research on heterogeneous graphs to construct topic evolution trees from the vocabulary of topics describing a domain (here we use the MeSH controlled vocabulary describing biomedical research). If a researcher is trying to understand a research topic outside their area of expertise, or a student is learning a new concept, this approach allows them to build on their prior knowledge and understand how a topic builds on concepts they may already understand.

METHOD
In this section, we use PubMed Central as a case study to investigate the generation of a query-centric Topic Evolution Tree (qTET).

Data Extraction
Scientific domains have developed detailed classification systems for categorizing research, and in this project we are focused on biomedical research which is classified in the PubMed database using the Medical Subject Headings (MeSH) controlled vocabulary, which as of 2014 contained 27,419 descriptors, which can be further constrained by 83 qualifiers. Prior research (Cheung, et al. 2012) has shown that some descriptors apply so broadly within PubMed that if relationships between research papers and concepts are based solely on descriptors, the network of relations is too large to provide guidance as to which terms were most relevant to a set of papers. However, the network connections can be filtered by leveraging both the MeSH qualifiers that constrain a concept, and metadata that identifies the descriptor/qualifier pair most important to a paper.

Context-rich Heterogeneous Network Construction
Prior studies have employed various methods to characterize citation networks, but the basic assumption is quite simple (Liu, et al., 2013): either one paper cites another paper, or one author cites another author; without considering the sentiment, reason, topic, or motivation for the citation. Most approaches assign equal importance (weight) to each reference cited in a paper, without considering that some references are central to the idea...
being presented in a paper while others are more tangential. This loss of information as to which references are most important to the topic of a particular paper limits the performance of retrieval or recommendation systems based on citation networks (Liu et al., 2014, Walker et al. 2007).

In this paper we first construct the heterogeneous graph shown in Figure 1 based on data from PubMed Central. This graph incorporates the relations described in Table 1. For each vertex in the graph, the edge weights for the same type of outgoing edge are normalized. For instance, the weight of the writtenby edge from paper, to author, is defined as

\[ w(p_i \rightarrow a_j) = \frac{1}{d(p_i \rightarrow A)} \]

where \( d(p_i \rightarrow A) \) is the total number of authors for paper \( p_i \). The weight of the co-author edge \( A \rightarrow A \) between two authors, as well as the citation edge \( P \rightarrow P \) between two papers are defined similarly. Since a paper can only be submitted to one venue, \( w(p \rightarrow v) = 1 \).

To estimate the contribution of each paper, venue, or author to a topic, the PageRank with Prior algorithm (White & Smyth, 2003) is used to calculate the paper, venue, and author importance for a given topic \( K \). The normalized topical PageRank authority score is used for the weights of \( K \rightarrow P \), \( K \rightarrow A \), and \( K \rightarrow V \). This step uses classical homogeneous graphs where each vertex is a paper, author, or venue and the citation relationship between vertices is utilized to calculate the PageRank authority scores. Each vertex is also characterized by a topic prior vector, i.e., for the paper graph, each paper has a topical prior distribution \( P(Z_k \mid p) \) that is the probability for each topic for that paper. For the author graph, the topical prior distribution is \( \sum_k P(Z_k \mid p) \) where \( p \), is published by the target author. The result of the PageRank (with prior) calculation is the topic authority vector for the paper, author, or venue. Note that topic \( j \) is contributed by paper \( p \), \( (K_j \rightarrow P) \) doesn’t necessarily mean paper \( p \), is relevant to topic \( j \), \( (P \rightarrow K) \). The relevance of a topic to a paper is based on whether the author has identified that topic as a major topic of the paper. In contrast, a paper could contribute to a topic even if the author did not specifically identify it as such. For example, some “HLA-DRB1” (a type of gene) papers can be important (e.g., contribute) to the topic “Type 1 diabetes”.

**Query-centric Topic Evolution via Meta-path on the Graph**

In this paper, we present a new method to characterize the evolution of a topic using heterogeneous graph mining that enables the construction of a query-centric Topic Evolution Tree (qTET) as shown in Figure 2. The root node of the qTET (\( k^* \) in Figure 2), is the topic a user wants to learn about; the inclusion of topics \( k_i \) through \( k_t \) as child nodes in the tree is based on their contribution to the topic represented by the root node.

The qTET in Figure 2 shows the evolution of a topic, and the probability associated with those evolutionary relationships is calculated based on the random walk of a meta-path on the heterogeneous graph, so \( P(k_i \rightarrow k_j \mid k^*) = RW_P(k_i \rightarrow k_j \mid k^*) \). A meta-path \( P \) is a path defined on the graph \( T_G = (\mathcal{N}, \mathcal{R}) \), where \( \mathcal{N} \) denotes the type of nodes in the graph and \( \mathcal{R} \) represents the possible relationships between nodes. For the heterogeneous graph in Figure 1,
the types of nodes in $\mathcal{N}$ are Paper, Keyword, Author, and Venue. Table 1 contains the possible relations between nodes. The meta-paths over the heterogeneous graph represent different types of evolutionary relationships between topics. An example meta-path is $K \rightarrow A \rightarrow K$ which denotes a meta-path between topics (keyword nodes) based on shared author(s).

We extend the classical meta-path as described above with additional constraints which we refer to as a restricted meta-path. The restricted meta-path allows for more sophisticated path instances; such as considering the extent to which a paper or its authors contributes to the user's query topic. We use $k^*$ to denote the user's query topic, and $k_1$ and $k_i$ to denote the start topic and end topics of the path ($k_i$ evolved from $k_1$). Following are two exemplar restricted meta-paths:

$$k^* \xleftarrow{w} P \xleftarrow{w} P \xrightarrow{r} k_i$$

$$\xleftarrow{con} k^* \xleftarrow{con} k^*$$

Topic $k^*$ evolves from topic $k_i$ via the citation path between papers, where both the cited and citing papers contribute to the initial query topic $k^*$. Example: if $k^* = \text{Gene X}$, $k_i = \text{Celiac Disease}$, and $k^* = \text{Type 1 Diabetes}$, then a path that matches this restricted meta-path would be Paper 1, focusing on Celiac Disease, cites Paper 2, which focused on Gene X, and both papers contributed to (but were not necessarily described by) the topic Type 1 Diabetes which was the user's query for generating the TET.

In the following example, instead of each paper contributing to the initial query topic $k^*$, each paper is written by an author who has respectively contributed to the topic $k^*$. This meta-path restriction loosens the above restriction on a paper contributing to a topic and instead looks at whether the respective authors of the papers have contributed significantly to the query topic. In calculating the scores for the edges in the TET, an author who has contributed more significantly to a topic (and is more likely to be an authority on the query topic) would be ranked higher in the results.

$$k^* \xleftarrow{w} P \xleftarrow{w} P \xrightarrow{r} k_i$$

$$\xleftarrow{con} k^* \xleftarrow{con} k^*$$

PRELIMINARY RESULTS

As a proof of concept, we implemented an experiment based on over 500,000 documents in PubMed Central (PMC) as of 2014. The XML for each document contained the full text, citation relationships, authors, publication venue, and identified the MeSH qualified descriptors that were the major topics of each paper. The distribution of articles for MeSH qualified descriptors (topics) had a long tail, with many qualified descriptors being used for only one article in PMC, so the top 3,000 qualified descriptors based on usage were included in the analysis. The distribution of articles for authors also exhibited a long tail, with many authors having a single article in PMC. The 155,050 authors with at least 3 papers on which they were the author or a co-author were included in the analysis. Pagerank calculations were run using the Jung framework for the contribution of a paper, author, or venue to each of the 3,000 qualified descriptors included in the analysis (9,000 calculations in total).

The data parsed from the PMC files for each article, along with the pagerank results for the contributions to each qualified descriptor, and data from the MeSH controlled...
vocabulary (when each term was added to the vocabulary) were loaded into a Neo4j graph database to construct the heterogeneous graph. Due to the number of papers, authors, and to a lesser extent, venues, the calculations for their contribution to each topic had a long-tailed distribution, so the top 500 papers, 100 authors, and 50 venues were included for each topic and the weights were then normalized. This resulted in the number of “contributed by” edges in the heterogeneous graph for papers being over 1.4 million, the number of edges for author contributions to a topic being 299,100, and the number of edges for venues contributions to a topic being 149,450. Due to multiple authors on papers, there were 720,790 “written by” edges and over 2 million co-author edges in the heterogeneous graph. The weights for the co-author edges were normalized based on the number of papers written with each co-author.

To test different restricted meta-path models, we constructed a framework that took a model as input expressed as JSON and used it for the random walk. In constructing the evolution tree, the maximum number of children a topic could evolve from, and the maximum depth of a branch in the evolution tree could be limited. For this preliminary analysis we set both to 5, so the maximum number of nodes in the evolution tree for any topic was 3,125. The output from the random walk is a topic evolution tree expressed in JSON that is then visualized in a web interface using Cytoscape.js as shown in Figure 3 (http://www.topicevolution.org/biomedical/TopicEvolution.html).

As a preliminary experiment, we compared 10 different meta-path models for 16 different MeSH qualified descriptors that were added to the MeSH controlled vocabulary over a wide span of years (1981-2014). For some of the earliest terms and one of the latest terms, the evolution trees were limited. For two terms we were most familiar with, we compared the two restricted meta-paths discussed earlier (based on restrictions to the contributions of the papers and authors to the search topic), with the unrestricted baseline meta-path: \( k \xleftarrow{\epsilon} p \xleftarrow{\epsilon} c \xrightarrow{r} k_i \).

The preliminary results showed that the unrestricted baseline meta-path performed poorly, often branching to unrelated concepts very early in the TET. In comparison, the TET’s based on the restricted meta-paths performed better as shown in Figure 3. Reviewing the pruning showed the baseline calculation was often selecting a large portion of the topics, many that were unrelated (e.g., seawater).

**CONCLUSIONS AND FUTURE WORK**

In this preliminary analysis we examined 10 meta-path models, and initial results showed topic evolution trees to be superior to approaches based solely on citations such as in the baseline meta-path. Although restricted meta-paths show promise in general for both newer topics and established topics that have been included in the vocabulary for over 20 years, further work is needed to validate the restricted meta-path results. Is there a meta-path model that is best for all topics or are there characteristics that make different models appropriate to different terms?

For this study we examined data from PubMed Central, and an additional avenue of exploration is to use the context of the paragraph in which a citation occurs to determine the motivation for a citation (since most articles are described by multiple MeSH terms). In addition to PubMed Central, the main PubMed database provides metadata on 27 million articles; an open question is whether that additional data would better refine the topic evolution tree.

**ACKNOWLEDGMENTS**

The authors thank Indiana University’s UITS Research Technologies for providing computational resources and Trevor Edelblute for invaluable assistance in running calculations on RT’s computing resources.

**REFERENCES**


