Information Extraction as Link Prediction: Using Curated Citation Networks to Improve Gene Detection

Andrew Arnold and William W. Cohen
Machine Learning Department, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
{aarnold, wcohen}@cs.cmu.edu

Abstract

In this paper we explore the usefulness of various types of publication-related metadata, such as citation networks and curated databases, for the task of identifying genes in academic biomedical publications. Specifically, we examine whether knowing something about which genes an author has previously written about, combined with information about previous coauthors and citations, can help us predict which new genes the author is likely to write about in the future. Framed in this way, the problem becomes one of predicting links between authors and genes in the publication network. We show that this solely social-network based link prediction technique outperforms various baselines, including those relying only on non-social biological information.

Introduction & related work

Social networks, in the form of bibliographies and citations, have long been an integral part of the scientific process. In this paper we examine how to leverage the information contained within these publication networks, along with information concerning the individual publications themselves and a user's history, to help predict which entities the user might be most interested in and thus intelligently guide his search.

Specifically, our application domain is the task of predicting which genes and proteins a biologist is likely to write about in the future (for the rest of the paper we will use the term 'gene' to refer both to the gene and gene product, or protein). We define a *citation network* as a graph in which *publications* and *authors* are represented as nodes, with bi-directional *authorship* edges linking authors and papers, and uni-directional *citation* edges linking papers to other papers (the direction of the edge denoting which paper is doing the citing and which is being cited). We can construct such a network from a given corpus of publications along with their lists of cited works. There exist many so called *curated* literature databases for biology in which publications are *tagged*, or manually labeled, with the genes with which they are concerned. We can use this metadata to introduce *gene* nodes to our enhanced citation network, which are bi-directionally linked to the papers in which they are tagged. Finally, we exploit a third source of data, namely biological domain expertise in the form of ontologies and databases of facts concerning these genes, to create *association* edges between genes which have been shown to relate to each other in various ways. We call the entire structure an *annotated citation network*.

While there has been extensive work on analyzing and exploiting the structure of networks such as the web and citation networks (Kleinberg 1999), most of the techniques used for identifying and extracting biological entities directly from publication text (Feldman et al. 2003; Murphy et al. 2004; Franzén et al. 2002; Bunescu et al. 2004; Shi and Campagne 2005) rely on performing named entity recognition on the text itself and ignore the underlying network structure entirely.

Data

We are lucky to have access to many sources of good data[1]:

- PubMed Central (PMC) contains full-text copies of over one million biological papers for which open-access has been granted.
- The Saccharomyces Genome Database(SGD) contains various types of information concerning the yeast organism *Saccharomyces cerevisiae*.
- The Gene Ontology (GO) describes the relationships between biological entities across numerous organisms.

From these we are able to extract the nodes and edges that make up our annotated citation network[2]:

Type	Name	Description	Number
Node	Paper		44,012
Node	Author		66,977
Node	Gene		5,816
Edge	Authorship	Author ↔ Paper	178,233
Edge	Mention	Paper ↔ Gene	160,621
Edge	Citation	Paper ↔ Paper	42,958
Edge	RelatesTo	Gene ↔ Gene	1,604

[1] pubmedcentral.nih.gov, yeastgenome.org, geneontology.org

[2] An on-line demo, including the network used for the experiments, can be found at http://yeast.ml.cmu.edu/nies/.

Methods

Given our graph representation, the first step is to pick a set of *query nodes* to which our predicted links will connect. We then perform a *random walk* out from the query node, simultaneously following each edge to the adjacent nodes with a probability proportional to the inverse of the total number of adjacent nodes (Cohen and Minkov 2006). We repeat this process a number of times, each time spreading our probability of being on any particular node, given we began on the query node. After each step in our walk we have a probability distribution over all the nodes of the graph, representing the likelihood of a walker, beginning at the query node(s) and randomly following outbound edges in the way described, of being on that particular node. We can then use this distribution to rank all the nodes, predicting that the nodes most likely to appear in the walk are also the nodes to which the query node(s) should most likely connect. In order to evaluate our predicted edges, we can hide certain instances of edges, perform a walk, and compare the predicted edges to the actual withheld ones.

Experiment

To evaluate our network model, we first divide our data into two sets:

- `Train` containing only *authors*, *papers*, *genes* and their respective relations published before 2008
- `Validation` containing new[3] relationships ($author \overset{Mentions}{\rightarrow} genes$) first published in 2008.

From this `Train` data we create a series of subgraphs (summarized in Figure 1), each emphasizing a different set of relationships between the nodes. By selectively removing edges of a certain type from the $FULL$ graph we were able to isolate the effects of these relations on the random walk and, ultimately, the predicted links. Specifically, we classify each graph into one of four groups and later use this categorization to asses the relative contribution of each edge type to the overall link prediction performance.

Baseline $UNIFORM$ is simply the chance of predicting a novel gene correctly given that you select a predicted gene uniformly at random from the universe of genes. Relatedly, ALL_PAPERS takes into account the distribution of genes across papers in the training graph. Thus its predictions are weighted by the number of times a gene was written about in the past. This model provides a more reasonable baseline. $AUTHORS$ considers the distribution of genes over all papers previously published by the author.

Social The social graphs are constructed of edges that convey information about the social interactions of authors, papers and genes. These include facts about which authors have written together, which papers have cited each other, and which genes have been mentioned in which papers.

Content In addition to social edges, some graphs also encode information regarding the biological content of the genes being published.

Protocol For our query nodes we select the subset of authors who have publications in both the `Train` and `Validation` set. We further create two other query author lists, FIRSTAUTHORS and LASTAUTHORS, restricted to those authors who appear as the first or last author, respectively, in their publications in the `Validation` set. The purpose of these lists of queries is to determine whether an author's position in a paper's list of authors has any impact in our ability to predict the genes he or she might be interested in.

Given these sets of graphs and query lists, we then query each author in each of our three lists, independently, against each subgraph in Figure 1. Each such (author, graph) query yields a ranked list of genes predicted for that author given that network representation. By comparing this list of predicted genes against the set of true genes from `Validation` we are able to calculate the performance of each (author, graph) pairing[4]. The resulting F1 metrics, broken down for each set of author positions, are summarized in Figure 2.

Querying with extra information Finally, we were interested in seeing what effect adding some limited information about an author's 2008 publications to our query would have on the quality of our predictions. This might occur, for instance, if we have the text of one of the author's new papers available and are able to perform basic information extraction to find at least one gene. We therefore also queried, together as a set, each author and the one new gene about which he published most in 2008 (see graph $FULL(AUTHOR{+}1\_GENE)$ in Figure1). These results are summarized, along with the others, in Figure 2, again broken down by author position.

Results

It is apparent from the results that, in almost all settings, querying based on the first author of a paper generates the best results, with querying by last author performing the worst. Tellingly, the only case in which the last author *is* most significant is in the $CITATIONS\_CITED$ model.

We notice that those models relying solely on the biological GO information relating genes to one another (**Content** graphs from Figure 1) perform significantly worse than any other model, and are in fact in the same range as the $UNIFORM$ model. Indeed, the $FULL$ model benefits from having the relations removed, as it is outperformed by the $FULL\_MINUS\_RELATED\_GENES$ model.

Some possible explanations for this are that scientists might not be driven to study genes which *have already been demonstrated* to be biologically related to one another. It is also possible that our methods for parsing and interpreting the GO information and extracting the relationships between

[3]We restrict our evaluation to genes about which the author has never previously published .

[4]Since the list of predicted genes can be quite long all evaluations are calculated only considering the top 20 predictions made.

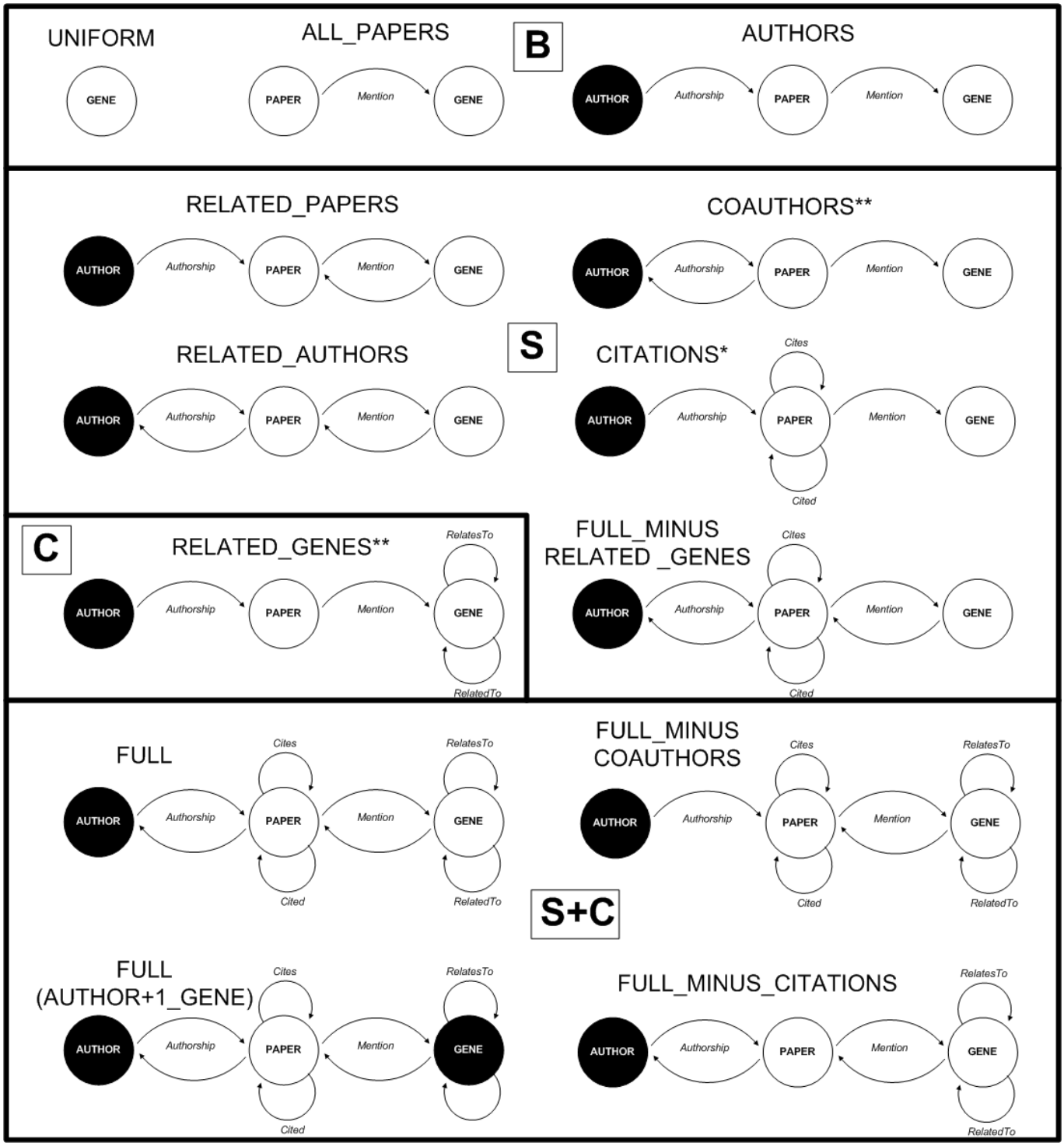

Figure 1: Subgraphs queried in the experiment, grouped by type: **B** for baselines, **S** for social networks, **C** for networks conveying biological content, and **S+C** for networks making use of both social and biological information. Shaded nodes represent the node(s) used as a query. **For graph $RELATED\_GENES$, which contains the two complimentary uni-directional *Relation* edges, we also performed experiments on the two subgraphs $RELATED\_GENES_{RelatesTo}$ and $RELATED\_GENES_{RelatedTo}$ which each contain only one direction of the *relation* edges. For graph $CITATIONS$, we similarly constructed subgraphs $CITATIONS_{Cites}$ and $CITATIONS_{Cited}$.

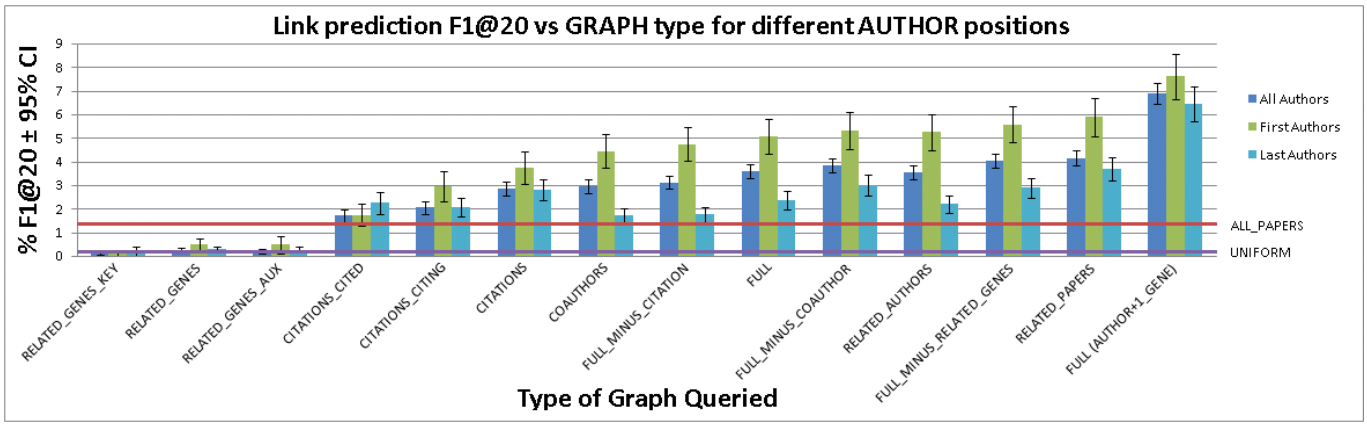

Figure 2: Mean percent F1 @20 of queries across graph types, broken down by author position, shown with error bars demarking the 95% confidence interval. Baselines $UNIFORM$ and ALL_PAPERS are also displayed.

genes may not be capturing the relevant information in the same way a trained biologist might be able to. R

In contrast, the models exploiting the **social** relationships in $CITATIONS$, $COAUTHORS$, $RELATED\_AUTHORS$ and $RELATED\_PAPERS$ all outperform the ALL_PAPERS baseline. While each of these social edge types is helpful on its own, their full combination is, perhaps counter-intuitively, not the best performing model. Indeed, while $FULL$ outperforms its constituent $CITATIONS$ and $COAUTHORS$ models, it nevertheless benefits slightly from having the *coauthor* edges removed (as in $FULL\_MINUS\_COAUTHOR$). This may be due to competition among the edges for the probability being distributed by our random walk.

The best performance of the single-author query models is achieved by the relatively simple, pure collaborative filtering $RELATED\_PAPERS$ model. This makes sense since, if other people are writing about the same genes as the author, they are more likely to share other common interests and thus would be the closest examples of what the author may eventually become interested in in the future.

The results for the $FULL(AUTHOR + 1\_GENE)$ model seem to indicate that adding a single known new gene to our author query of the $FULL$ model improves our prediction performance by almost 50%, and significantly outperforms the best single-author query model, $RELATED\_PAPERS$. This is a promising result, as it suggests that the information contained in our network representation can be combined with other sources of data (e.g. gleaned from performing information extraction on papers' text) to achieve even better results than either method alone.

Conclusions & future work

In this paper we have introduced a new graph-based annotated citation network model to represent various sources of information regarding publications in the biological domain. We have shown that this network representation alone, without any features drawn from text, is able to outperform competitive baselines. Using extensive ablation studies we have investigated the relative impact of each of the different types of information encoded in the network, showing that social knowledge often trumps biological content, and demonstrated a powerful tool for both combining and isolating disparate sources of information. We have further shown that, in the domain of Saccharomyces research from which our corpus was drawn, knowing who the first author of a paper is tends to be more informative than knowing who the last author is (contrary to some conventional wisdom). Finally, we have shown that, despite performing well on its own, our network representation can easily be further enhanced by including in the query set other sources of knowledge about a prediction subject gleaned from separate techniques, such as information extraction and document classification.

References

Bunescu, R.; Ge, R.; Kate, R.; Marcotte, E.; Mooney, R.; Ramani, A.; and Wong, Y. 2004. Comparative experiments on learning information extractors for proteins and their interactions. In *Journal of AI in Medicine*.

Cohen, W. W., and Minkov, E. 2006. A graph-search framework for associating gene identifiers with documents. *BMC Bioinformatics* 7(440).

Feldman, R.; Regev, Y.; Finkelstein-Landau, M.; Hurvitz, E.; and Kogan, B. 2003. Mining the biomedical literature using semantic analysis. *Biosilico* 1(2):69–80.

Franzén, K.; Eriksson, G.; Olsson, F.; Asker, L.; Lidn, P.; and Cöster, J. 2002. Protein names and how to find them. In *International Journal of Medical Informatics*.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. In *JACM*.

Murphy, R. F.; Kou, Z.; Hua, J.; Joffe, M.; and Cohen, W. W. 2004. Extracting and structuring subcellular location information from on-line journal articles: The subcellular location image finder. In *KSCE*.

Shi, L., and Campagne, F. 2005. Building a protein name dictionary from full text: a machine learning term extraction approach. *BMC Bioinformatics* 6(88).