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

Andrew Arnold and William W. Cohen
{aarnold, wcohen}@cs.cmu.edu

Machine Learning Department, Carnegie Mellon University

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.

1 Introduction & Related Work

Social networks, in the form of bibliographies and citations, have long been an integral part of the scientific process. Most scientists begin their exploration of a new problem with an intense investigation of the relevant literature. In a new or small field, for which the universe of such citations is relatively small, both a broad and deep search is manageable. As the size of the set of related papers grows, however, a researcher's time and attention can easily become overwhelmed. While the Internet has provided scientists with new tools for performing these literature reviews more quickly and precisely, it is usually left up to the user to guide the search themselves. In other words, one has to know what she is looking for. At the same time the space of accessible, and possibly relevant, papers has increased even more swiftly, leaving many valuable publications undiscovered. This is the problem we address in this paper: 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*.

Although academics have long recognized and investigated the importance of such networks, their investigations have often been focused on historical [1], summary, or explanatory purposes [2–5]. While other work has been concerned with understanding how influence develops and flows through these networks [6], we instead focus on the problem of link prediction [7, 8]. *Link prediction* is the problem of predicting which nodes in a graph, currently unlinked, ”should” be linked to each other, where ”should” is defined in some application-specific way. This may be useful to know if a graph is changing over time (as in citation networks when new papers are published), or if certain edges may be hidden from observation (as in detecting insider trading cabals). In our setting, we seek to discover edges between authors and genes, indicating genes about which an author has yet to write, but which he may be interested in.

While there has been extensive work on analyzing and exploiting the structure of networks such as the web and citation networks [9, 10], most of the techniques used for identifying and extracting biological entities directly from publication text [11–16] and curated databases [17] rely on performing named entity recognition on the text itself [18] and ignore the underlying network structure entirely. While these techniques perform well given a paper to analyze, they are impossible to use when such text is unavailable, as in our link prediction task.

In the following sections, respectively, we discuss the topology of our annotated citation network, along with describing the data sources from which the network was constructed. We then introduce *random walks*, the technique used for calculating the proximity of nodes in our graph, thus suggesting plausible novel links between authors and genes. Finally, we describe an extensive set of ablation studies performed to assess the relative importance of each type of edge, or *relation*, in our model and discuss the results, concluding with a view towards a future model combining network and text information.

2 Data

We are lucky to have access to many sources of good data[1] from which we are able to extract the nodes and edges that make up our annotated citation network[2]:

- PubMed Central (PMC) contains full-text copies of over one million biological papers for which open-access has been granted.

[1] http://pubmedcentral.nih.gov, http://yeastgenome.org, http://geneontology.org

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

– 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.

Nodes	
Name	**Number**
Paper	44,012
Author	66,977
Gene	5,816

Edges		
Name	**Description**	**Number**
Authorship	Author ↔ Paper	178,233
Mention	Paper ↔ Gene	160,621
Citation	Paper ↔ Paper	42,958
RelatesTo	Gene ↔ Gene	1,604

3 Methods

Now that we have a representation of the data as a graph, we are ready to begin the calculation of our link predictions. The first step is to pick a node, or set of nodes, in the graph to which our predicted links will connect. These are our *query nodes*. 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 [19]. 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. If there are multiple nodes in the query set, we perform our walk simultaneously from each one. 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. Under the right conditions, after enough steps this distribution will converge. 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 practice, the same results can be achieved by multiplying the adjacency matrix of the graph by itself. Each such multiplication represents one complete step in the walk.

We can adjust the adjacency matrix (and thus the graph) by selectively hiding, or removing, certain *types* of edges. For instance, if we want to isolate the influence of citations on our walk, we can remove all the citation edges from the graph, perform a walk, and compare the results to a walk performed over the full graph.

Likewise, 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. For example, if we have all of an author's publications and their associated gene mention data for the years 2007 and 2008, we can remove the links between the author and the genes he mentioned in 2008 (along with all other edges gleaned from 2008 data), perform a walk, and then see how many of those withheld gene-mention edges were correctly predicted. Since this evaluation is a comparison between one unranked set (the true edges) and another ranked list (the predicted edges) we can use the standard information retrieval metrics of precision, recall and F1.

4 Experiment & Results

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

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

From this `Train` data we create a series of subgraphs, each emphasizing a different set of relationships between the nodes. These subgraphs are summarized in Figure 1. 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 assess the relative contribution of each edge type to the overall link prediction performance.

4.1 Baseline

The baseline graphs are $UNIFORM$, ALL_PAPERS and $AUTHORS. UNIFORM$ and ALL_PAPERS do not depend on the *author* node. $UNIFORM$, as its name implies, 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. Since there are 5,816 gene names, and on average each author in our query set writes about 6.7 new genes in 2008, the chance of randomly guessing one of these correctly is $6.7/5816 = .12\%$. Using these values we can extrapolate this model's expected precision, recall and F1. Relatedly, ALL_PAPERS, while also independent of authors, nevertheless 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. While this type of model may help recover previously published genes, it may not do as well identifying new genes.

4.2 Social

The social graphs ($RELATED\_PAPERS$, $RELATED\_AUTHORS$, $COAUTHORS$, $FULL\_MINUS\_RELATED\_GENES$ and $CITATIONS$) 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.

[3] We restrict our evaluation to genes about which the author has never previously published (even though an author may publish about them again in 2008), since realistically, these predictions would be of no value to an author who is already familiar with his own previous publications.

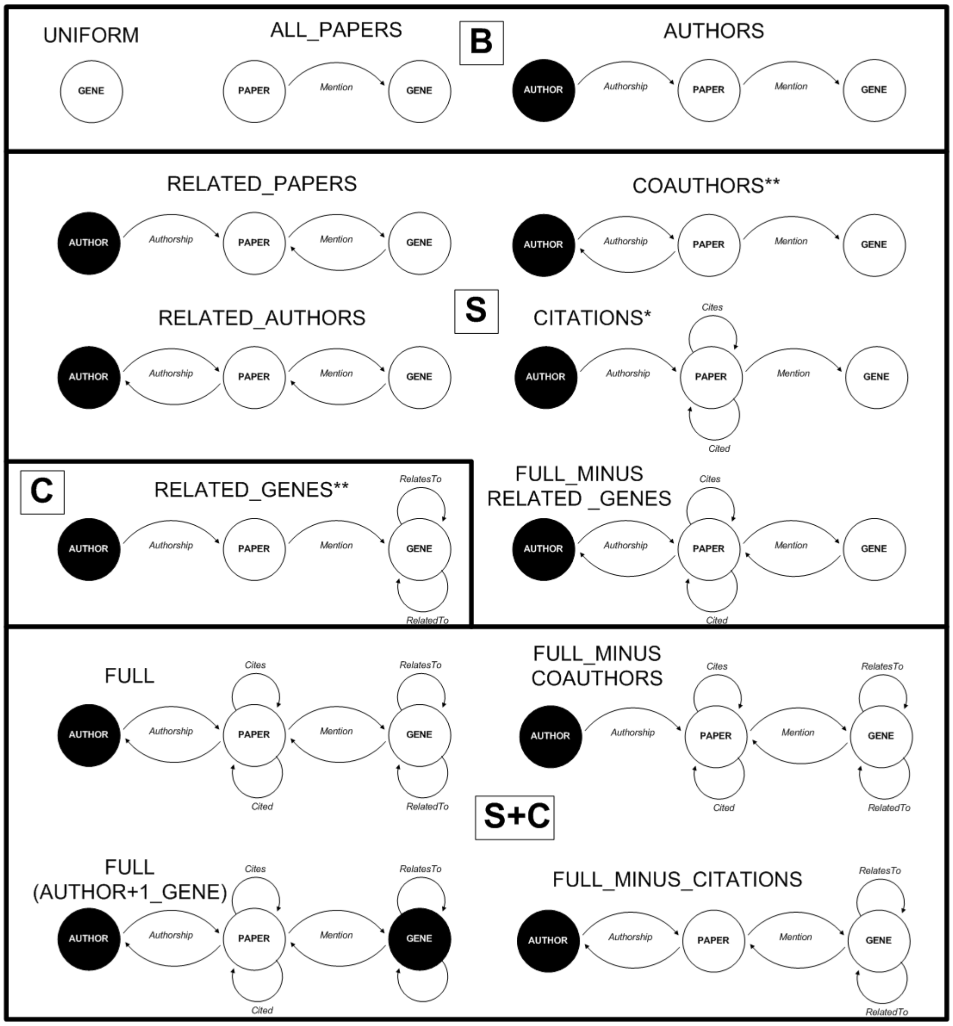

Fig. 1. Subgraphs queried in the experiment, grouped by type: **B** for baselines, **S** for social networks, **C** for networks conveying biological content, and **S+C** for networks making use of both social and biological information. Shaded nodes represent the node(s) used as a query. **For graph $RELATED\_GENES$, 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 $CITATIONS_{Cited}$.

4.3 Content

In addition to social edges, some graphs also encode information regarding the biological content of the genes being published. The graph $RELATED\_GENES$ models only this biological content, while $FULL\_MINUS\_COAUTHORS$, $FULL\_MINUS\_CITATIONS$, $FULL$ and $FULL(AUTHOR + 1\_GENE)$ all contain edges representing both social and biological content.

4.4 Protocol

For our query nodes we select the subset of authors who have publications in both the `Train` and `Validation` set. To make sure we have fresh, relevant publications for these query authors, and to minimize the impact of possible ambiguous name collision, we further restrict the query author list to only those authors who have publications in both 2007 and 2008. This yields a query list, ALLAUTHORS, with a total of 2,322 authors, each to be queried independently, one at a time. We further create two other query author lists, FIRSTAUTHORS and LASTAUTHORS containing 544 and 786 authors respectively, 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. Since the list of predicted genes is sometimes quite long (since it is a distribution over all genes in the walk), we set a threshold and all evaluations are calculated only considering the top 20 predictions made. These resulting precision, recall, F1 and MAP metrics, broken down for each set of author positions, are summarized in Figure 2 respectively.

4.5 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. The question is, can we leverage this single, perhaps easy to identify gene, to improve our chances of predicting or identifying other undiscovered new genes? To answer this question, in addition to querying each author in isolation, we 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.

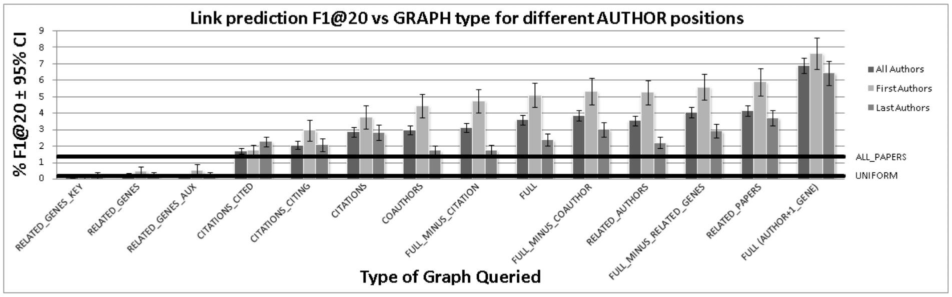

Fig. 2. Mean percent F1 @20 of queries across graph types, broken down by author position, shown with error bars demarking the 95% confidence interval. Baselines $UNIFORM$ and ALL_PAPERS are also displayed.

4.6 Results

Using Figures 1 and 2 as guides, we turn now to an analysis of the effects different edge types have on our ability to successfully predict new genes[4]. We should first explain the absence of results for the $AUTHORS$ graph, and the lines for $UNIFORM$ and ALL_PAPERS in Figure 2. Since these baselines do not depend on the query, they are constant across models and are thus displayed as horizontal lines across the charts in Figure 2. $AUTHORS$ is missing because it is only able to discover genes that have *already* been written about by the query authors in the training graph. Since our evaluation metrics only count the prediction of novel genes, $AUTHORS$'s performance is necessarily zero.

Given these baselines, let us next consider the role of author position on prediction performance. 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. This seems to suggest that knowing the first author of a paper is more informative than knowing who the last author was in terms of predicting which genes that paper may be concerned with. Depending on the specifics of one's own discipline, this may be surprising. For example, in computer science it is often customary for an advisor, lab director or principal investigator to be listed as the last author. One might assume that the subject of that lab's study would be most highly correlated with this final position author, but the evidence here seems to suggest otherwise. Tellingly, the only case in which the last author *is* most significant is in the $CITATIONS\_CITED$ model. Recall that in this model edges from cited papers to their citing papers are present. These results may suggest that in this model, knowing the last author of the paper actually is more valuable.

Given that in most cases the models queried using first authors performed the best, the columns of Figure 2 have been positioned in order of increasing first author F1 performance, and all subsequent comparisons are made with respect to the first author queries, unless otherwise stated. Thus we notice that those models relying solely on the biological GO information relating genes to one another (**Content** graphs from Figure

[4] A summery of the claims made and their associated statistical tests are summarized in Table 1.

Claim	**Statistical test**
Last author is most significant in $CITATIONS\_CITED$	80% confidence intervals
Content graphs perform worse than any other model	Wilcoxon signed rank (p $<$ .01)
Content graphs are in the same range as $UNIFORM$	Inside 95% confidence intervals
Removing $RELATED\_GENES$ improves $FULL$	Wilcoxon signed rank (p $<$ .01)
Social graphs outperform ALL_PAPERS	Outside 95% confidence intervals
$FULL$ outperforms $CITATIONS$ and $COAUTHORS$	Wilcoxon signed rank (p $<$ .01)
$FULL$ benefits from having *coauthor* edges removed	Wilcoxon signed rank (p $<$ .15)
$RELATED\_PAPERS$ is best single-author query model	Wilcoxon signed rank (p $<$ .10)
$FULL(AUTHOR + 1\_GENE)$ performs best	Paired sign (p $<$ .02)

Table 1. A summary of the claims made and the statistical tests used to support those claims.

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.

There are a few possible explanations for why these content-based biological edges might be hurting performance. First, scientists might not be driven to study genes which *have already been demonstrated* to be biologically related to one another. Since we are necessarily using biological facts already discovered, we may be behind the wave of new investigation. Second, these new investigations, some of them biologically motivated, might not always turn out conclusively or successfully. This would likewise lead to the genes being studied in this way lying outside the scope of our biological content. Finally, it is possible that our methods for parsing and interpreting the GO information and extracting the relationships between genes may not be capturing the relevant information in the same way a trained biologist might be able to. Relatedly, the ontologies themselves might be designed more for summarizing the current state of knowledge, rather than suggesting promising areas of pursuit.

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 more paths there are out of a node, the less likely the walker is to follow any given one. Thus, by removing the (many) coauthorship edges from the $FULL$ graph, we allow the walk to reach a better solution more quickly.

Interestingly, the best performance of the single-author query models is achieved by the relatively simple, pure collaborative filtering $RELATED\_PAPERS$ model [20]. Explained in words, this social model predicts that authors are likely to write about genes that co-occur with an author's previously studied genes in other people's papers.

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.

Finally we examine the question of whether having not only a known author to query, but also one of this author's new genes, aids in prediction. The results for the $FULL(AUTHOR + 1\_GENE)$ model[5] seem to indicate that the answer is yes. 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, $RELATED\_PAPERS$, as well. This is a promising result, as it suggests that the information contained in our network representation can be combined with other sources of data (gleaned from performing information extraction on papers' text, for example) to achieve even better results than either method alone.

5 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.

We plan to extend this work by incorporating the results of these social network models into standard information extraction techniques. Since the end result of our link prediction algorithm is a distribution over nodes, one simple way to do this would be to use that distribution as a prior for a probabilistic information extraction methods. We also see value in incorporating a temporal dimension to our network. In our current model all edges are walked upon with equal probability, regardless of the temporal distance between the two connected nodes. We might do better by taking this time distance into account: for example, coauthorship on a paper 20 years ago may carry less weight than a collaboration just a few years ago.

References

1. Garfield, E., Sher, I., Torpie, R.: The Use of Citation Data in Writing the History of Science. The Institute for Scientific Information (1964)

[5] During evaluation the queried new gene is added to the set of previously observed genes and thus does not count towards precision or recall.

2. Erosheva, E., Fienberg, S., Lafferty, J.: Mixed membership models of scientific publications. PNAS **101**(21) (2004)
3. Liu, X., Bollen, J., Nelson, M., de Sompel, H.V.: Co-authorship networks in the digital library research community. In: Information Processing and Management,. (2005)
4. Cardillo, A., Scellato, S., Latora, V.: A topological analysis of scientific coauthorship networks. In: Physica A: Statistical Mechanics and its Applications. (2006)
5. Leicht, E.A., Clarkson, G., Shedden, K., Newman, M.E.J.: Large-scale structure of time evolving citation networks. Eur. Phys. J. B **59** (2007) 75–83
6. Dietz, L., Bickel, S., Scheffer, T.: Unsupervised prediction of citation influences. In: ICML. (2007)
7. Cohn, D., Hofmann, T.: The missing link: A probabilistic model of document content and hypertext connectivity. In: NIPS. (2001)
8. Liben-Nowell, D., Kleinberg., J.: The link prediction problem for social networks. In: CIKM. (2003)
9. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. In: JACM. (1999)
10. Kleinberg, J.M., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.S.: The web as a graph: Measurements, models and methods. In: Lecture Notes in Computer Science. (1999)
11. Cohen, A.M., Hersh, W.R.: A survey of current work in biomedical text mining. Briefings in Bioinformatics **6** (2005) 57–71
12. Feldman, R., Regev, Y., Finkelstein-Landau, M., Hurvitz, E., Kogan, B.: Mining the biomedical literature using semantic analysis. Biosilico **1**(2) (2003) 69–80
13. Murphy, R.F., Kou, Z., Hua, J., Joffe, M., Cohen, W.W.: Extracting and structuring subcellular location information from on-line journal articles: The subcellular location image finder. In: KSCE. (2004)
14. Franzén, K., Eriksson, G., Olsson, F., Asker, L., Lidn, P., Cöster, J.: Protein names and how to find them. In: International Journal of Medical Informatics. (2002)
15. Bunescu, R., Ge, R., Kate, R., Marcotte, E., Mooney, R., Ramani, A., Wong, Y.: Comparative experiments on learning information extractors for proteins and their interactions. In: Journal of AI in Medicine. (2004)
16. Shi, L., Campagne, F.: Building a protein name dictionary from full text: a machine learning term extraction approach. BMC Bioinformatics **6**(88) (2005)
17. Wang, R.C., Tomasic, A., Frederking, R.E., Cohen, W.W.: Learning to extract gene-protein names from weakly-labeled text. In: CMU SCS Technical Report Series (CMU-LTI-08-04). (2006)
18. Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. (1999)
19. Cohen, W.W., Minkov, E.: A graph-search framework for associating gene identifiers with documents. BMC Bioinformatics **7**(440) (2006)
20. Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Communications of the ACM **35**(12) (1992) 6170