Abstract - Knowledge Discovery algorithms generate alerts to interesting activity by finding partial matches to user-specified patterns of interest. Several statistical models can be used to generate pattern based on training data. However, when training sets are not readily available it is still possible to construct patterns from domain knowledge. To test the feasibility of such an approach we examine a database of publications in biological sciences, BioBase, and attempt to predict whether or not two researchers who did not coauthor a paper in 1998-2002 will coauthor a paper in 2003. We constructed a set of four distinct patterns depicting scenarios in which a new coauthoring relationship might emerge. In each of the scenarios, we needed to identify individuals playing different roles within research groups, and use the number of publications of individuals to predict those roles. We also use the 1998-2002 coauthor relationships to provide evidence of the associated research groups and collaborations. Our results in a test database show that this approach is feasible and competitive when compared to others that rely on more extensive statistical modeling.

Keywords: modeling, pattern based fusion.

1 Overview

Knowledge discovery and dissemination (KDD) technologies are emerging as new and powerful tools for the mining of relational data in applications as diverse as fraud detection, counterterrorism, and bioinformatics. These technologies include but are not limited to: information extraction, entity resolution and characterization, behavior prediction, organizational structure and threat activity detection, and monitoring for anomalous activities. While many of these technologies are beginning to prove themselves in selected applications with real data, we still need more tests and evaluations to assess how well they actually address, individually and collectively, the challenges of very large real datasets with low signal-to-noise and limited or no ground truth.

We have developed a tool that is a logic-based modeling approach that attempts to match user-authored behavior patterns with available, usually incomplete evidence. Unlike statistical, learning-based approaches used by other teams, the tool does not require training data, an important advantage in situations where ground truth is not available.

1.1 Problem Description

We focused initially in testing the potential of our approach on a prediction task. In particular we attempt to predict whether two authors that have not published together in the years 1998-2002 will publish together in 2003. The availability of databases with the required information made this problem interesting. We used BioBase, a collection of publication records from the biological sciences from 1998 to 2002. Each record in BioBase corresponds to a single journal publication, and incompletely contains authors, title, journal, keywords, and other possible information about the article. Because of the size of BioBase, we only attempt to provide predictions for a restricted set of potential co-authors, rather than on all possible co-authorship pairs. The experiment is carried out on 500 persons of interest, each of which has an associated list of 100 potential 2003 co-authors. For each of the 500*100=50,000 potential co-author pairs, we will provide a probability that the given pair will co-author one or more papers in 2003.

1.2 Technology Description

While manual link analysis methods for associating evidence to analyst-derived patterns can be usefully applied to small data sets, they break down when data sets become large or densely connected, or when relevant clues are so widely scattered that they cannot be conveniently localized on a single visual display. These challenges become all the more difficult when the task is to detect a terrorist threat before an attack occurs, as opposed to investigating after the fact since associated evidence may be less complete and harder to fuse. This suggests a need for new relational data mining methods to help automate link analysis. Such link discovery systems include our tool, developed by BAE Systems.

Our tool is a link detection system that takes in a threat pattern—a pattern of interesting behavior or activity—and observed data containing trace observations about threat activity, and outputs threat hypotheses with inferred actors and events. We use a Prolog-based frame system to represent threat patterns and enforce temporal and equality constraints among pattern slots. Based on rules involving uniquely identifying slots in the pattern, we trigger an initial set of threat hypotheses, and then refines these hypotheses by generating queries for unknown slots from constraints involving known slots.
To evaluate hypotheses, we score each local (i.e., single event) hypothesis using a probabilistic model in order to create a consistent, high-value global hypothesis by pruning conflicting lower scoring hypotheses.

Figure 1. We generate and evaluate hypotheses using structured evidence and pre-defined patterns, and generates reports for analysts based on the highest quality hypotheses.

Figure 1 shows our system’s functionality. The primary inputs are:

- Entities and links (including relationships and communications) in the structured database of evidence
- A pattern of interest describing relationships between key events and roles in the activity of interest.

We use these inputs to generate threat hypotheses through two main stages: generation and evaluation. We combine two key approaches—abductive reasoning to identify partial instantiations of threat pattern activity occurring in link data for hypothesis generation, and probabilistic evaluation to rank the pattern matches according to the most likely or most threatening activities for hypothesis evaluation. After ranking the threat events, we alert the analysts to the most threatening of the events, and can provide information to those in charge of collection about the most useful, but currently unknown, information. Further, we can use the threat hypotheses describing the individual predicted threat events—captured in the local hypotheses—to construct a global hypothesis. A global hypothesis is highest quality set of individual threat hypotheses that are non-contradictory; a global hypothesis provides analysts with the most likely global threat assessment given the available data. For more information see [1].

2 Description of Application

Our approach to predicting new coauthor relationships is twofold. We began our analysis by establishing a baseline for coauthorship occurrence, empirically determining the probability that two individuals will coauthor a paper. After establishing this baseline coauthorship score, we applied our tool to identify pairs of individuals that are likely to coauthor a paper based on using the known coauthorships in 1998-2002. We then refine our baseline coauthorship scores depending on whether the tool predicts two individuals are likely to coauthor a paper under one of our predefined scenarios.

2.1 Establishing baseline performance

We focus our attention on the 3050 distinct authors that appear in the 50,000 candidate coauthorship pairs; we believe that statistics calculated for this set of individuals are more relevant than statistics calculated for the BioBase population at-large. We will assume a prior probability of collaboration among the pairs to be 5%. This assumption is based on rough estimates and could significantly impact the performance of the algorithm. Lower prevalences make the identification problem harder, so we biased our estimates toward a smaller prevalence.

One of the best indicators as to whether two individuals will coauthor a paper is the existence of a common coauthor in a preceding year. That is, if a third person exists that has coauthored papers with both of the individuals, the two individuals are more likely to coauthor a paper. Using a sample of the BioBase data, we calculated that 10% of the possible new coauthor pairs of individuals who shared a common coauthor in the publications from 1998-2001, coauthored a paper in 2002 and only 2% otherwise. This simple test provides already some resolution over the presumed prior of 5%.

2.2 Activity patterns

We generate alerts to interesting activity by finding partial matches to user-specified patterns of interest. Since the objective of our experiment is to predict whether or not two people who did not coauthor a paper in 1998-2002 will coauthor a paper in 2003, we constructed a set of four distinct patterns depicting scenarios in which a new coauthoring relationship might emerge. In each of the scenarios (explained below), we need to identify individuals playing different roles within research groups, and use the number of publications of individuals to predict those roles. We also use the 1998-2002 coauthor relationships to provide evidence of the associated research groups and collaborations.

The first scenario involves the transfer of a junior researcher to a new research group. Such scenario often occurs when a student completes PhD or post doctorate research with an advisor and moves into a new research group as part of a new professorship or post doctorate position. When a new researcher joins a research group, s/he is likely to coauthor papers with other members of the new research group.

In the first scenario, depicted in Figure 2, we look for evidence of two research groups and their leaders. Leaders of research labs are subject matter authorities, and typically have a large number of publications, and so we look for two authors with large numbers of publications. (Here, we consider “large” to be 30 publications.) Next, we list the members of their research labs by listing their associated coauthors. We then identify an individual (JuniorResearcher1 in Figure 2) that is likely a transfer from one research lab to the other. Evidence for a transfer is that the individual publishes articles prior to 2002 with the leader of the first lab, but in 2002 publishes instead with the leader of the second lab—a newly formed relationship, since the individual has not published with members of the second research lab prior to 2002. When we detect such a pattern, we
predict that the new member of the second research lab will coauthor a paper with another junior member of the lab.

![Figure 2](image.png)

Figure 2. A junior researcher is likely to coauthor a paper with other members of a new research group.

In the second scenario, we identify two research labs that may be working together as part of a larger research project. Such lab partnerships emerge from larger, cross-disciplinary research funding, or when the interests of two different labs intersect in an area that both groups wish to explore jointly. When research labs begin working together, junior researchers from the two labs are more likely to coauthor papers.

To identify occurrences of the second scenario, depicted in Figure 3, we again find research leaders and their associated research lab members by looking for individuals with a large number of publications and their coauthors. When we identify two potential research labs, we look for evidence of collaboration between the two labs. Evidence for this collaboration includes coauthorships tying the research leaders to each other and to junior members of the opposing research lab. When we identify this combination of evidence, we predict that pairs of junior members from the two labs, such as JuniorResearcher1 from ResearchLab1 and JuniorResearcherK from ResearchLab2 in Figure 3, are likely to coauthor a paper.

![Figure 3](image.png)

Figure 3. When two research groups team together to work on a larger research effort, members of the labs are likely to coauthor papers.

For the third scenario, we look for research leaders that exhibit working relationships or other affiliations that would encourage them to coauthor papers. Once the research leaders from two labs work together, there is an increased likelihood that the leaders will reach out to the junior researchers in each other’s labs; if we see evidence that such reaching out occurs, we can predict that further instances of this kind of collaboration and corresponding coauthorship will occur.

When looking for instances of the third scenario, depicted in Figure 4, we once again look for research leaders and their associated labs by identifying individuals with a large number of publications and their associated coauthors. We then look for instances in which two research leaders directly coauthor a paper. We further refine the scenario by requiring at least one of the research leaders to coauthor a paper with at least one of the junior researchers from the opposing lab. Once we identify such instance, we predict that other junior members from the labs with coauthor papers with the research leader from the opposing lab.

![Figure 4](image.png)

Figure 4. When senior researchers have a working partnership, they may coauthor papers with junior researchers from each other’s research groups.

In the fourth scenario, we consider the simplest pattern, which is that people belonging to the same research group are likely to publish articles with one another. The nature of research labs leads subsets of lab members to publish together, but all lab members are not part of every publication. Consequently, there may be two lab members that have published papers with other members of the lab prior to 2003, but have not yet published together. When we identify such possible pair, we predict that they will publish together in 2003.

In the fourth scenario, depicted in Figure 5, we do not need to distinguish junior or senior research lab members; we simply need to find a group of collaborating authors. We look for a pair of individuals (Researcher1, Researcher2 in Figure 5) that has at least two common coauthors, and predict that all the involved individuals are part of the same research lab. To further strengthen the evidence that we have detected a research lab, we also require that the two common coauthors themselves coauthor a paper with each other. We then predict that the original pair will coauthor a paper in 2003.
Figure 5. Members of a common research group are likely to coauthor papers.

2.3 Generating Coauthorship Probability Scores

Our analysis showed that scenario 4 was the more reliable pattern. The predictive power of the other scenarios, when discounting the pairs already generated by scenario 4, was significantly reduced.

We thus generate coauthorship probability predictions using the following rules:

- If a candidate pair of potential coauthors (X,Y):
  - Is predicted to coauthor a paper by scenario 4 (i.e., X and Y share a common coauthor in 1998-2002), we assign a coauthorship probability score of 0.19, the precision of the test for the estimated prevalence, for using scenario 4 to predict coauthorship.
  - Else, if (X,Y) share a common coauthor, we assign the probability 0.10, the true negative rate for using common coauthorships as a predictor for future coauthorships.
  - Else, we assign the default probability of 0.02, the precision of the prediction of collaboration given that there is no common coauthor.

In other domains, these scores could be derived from subject matter experts instead of from empirical measures.

We also used this restricted set of data to test our CADRE application. For each scenario described above, we report in Table 1 the true positive rate and false positive rate of the associated coauthor predictions we make. Based on our test sample, we believe that the simplest pattern—identifying members of a common research group that have not yet coauthored a paper—is the best predictor of future coauthorship.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>True positive rate</th>
<th>True negative rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Junior researcher joins new lab</td>
<td>0.248</td>
<td>0.602</td>
</tr>
<tr>
<td>2: Research labs collaborate</td>
<td>0.130</td>
<td>0.960</td>
</tr>
<tr>
<td>3: Research leaders reach out to each others’ labs</td>
<td>0.058</td>
<td>0.996</td>
</tr>
<tr>
<td>4: Members of a common research lab coauthor papers</td>
<td>0.315</td>
<td>0.993</td>
</tr>
</tbody>
</table>

Two common tools for analyzing the performance of binary predictors are the Receiver Operator Curve (ROC) and the area under the ROC denoted AOC. The ROC is obtained by plotting the True positive rate (Tpr, the ratio of correct positive predictions to the actual number of positive cases) versus the False Positive Rate (Fpr, the ratio of incorrect positive predictions to the number of actual negative cases), for all possible values of a decision threshold. The AOC is the area under the ROC curve. A process with no prediction value will have the diagonal as ROC and 0.5 as AOC. A perfect classifier will have the left and top edges of the plot as ROC and 1 as AOC. Figure 6 shows the AOC for the experiment we carried out. The corresponding AOC is 0.8.

Figure 6: Receiver Operator Curve (ROC) for the experiment results

We compared our results to those obtained by other researchers tackling the same problem using conventional statistics tools. Our approach ranked favorably among them trailing one and tying with a second one out of a total of seven.

3.2 Notes

1. Many of the patterns of collaboration within the research community, such as evolving research interests, topic trends, and other research career evolutions that we might be able to capture with our tool actually occur in longer time spans than our data sets supported. We believe that our approach would have benefited from more years’ worth of data.

2. Many technologies can be tuned with better knowledge of the domain. We were able to draw upon our own research experience to define some example scenarios, but would be able to provide better patterns and corresponding results with increased knowledge of the domain. For example, the following background information about publications in the biological sciences may have been useful to us and to others:

- Are research funding sources typically listed as authors?
- Is the order of author names significant for a publication?
- What conventions are followed for topic identification or for identifying affiliations?
4. Identifying common training data is critical if the results of the technologies are to be compared. If BAE Systems, for example, identified a subset of the data for learning models, and another technology developer used a different set of data for the same purpose, the results may not be comparable.

4 Suggestions for Future Work

We would like to recommend that a future challenge problem involving BioBase or a similar dataset focus on rare event detection capabilities. Identification of the appropriate rare event may require further analysis of BioBase or input from a researcher in the biological sciences, but example rare events may include:

- Predicting that a researcher will receive an award or recognition for this research, or
- Predicting which paper or papers received honors as “Best Submission” or “Best Student Paper” at a conference.

If obtainable, we would like to encourage the use of data with a broader range of relationships. BioBase, for example, could be supplemented with other relationship sets to include links between researchers and organizations, researchers and conferences attended, and researchers and universities or research companies, assuming the cross-database entity resolution is completed. (In fact, the cross-database entity resolution itself may prove to be an interesting challenge problem.)

We would also like to combine prediction and entity resolution to make further inroads into a test that will accurately represent the challenges of multi-source data fusion. The prediction results could be improved by cleaning the datasets according to entity resolution and disambiguation tools or modify the patterns to account for instances in which an author appears under more that one name in the dataset.

References