Scientific Discovery as Link Prediction in Influence and Citation Graphs

Fan Luo Marco Valenzuela-Escárcega Gus Hahn-Powell Mihai Surdeanu
University of Arizona, Tucson, AZ, USA
{fanluo, marcov, hahnpowell, msurdeanu}@email.arizona.edu

Abstract

We introduce a machine learning approach for the identification of "white spaces" in scientific knowledge. Our approach addresses this task as link prediction over a graph that contains over 2M influence statements such as "CTCF activates FOXA1", which were automatically extracted using open-domain machine reading. We model this prediction task using graph-based features extracted from the above influence graph, as well as from a citation graph that captures scientific communities. We evaluated the proposed approach through backtesting. Although the data is heavily unbalanced (50 times more negative examples than positives), our approach predicts which influence links will be discovered in the "near future" with a F1 score of 27 points, and a mean average precision of 68%.

1 Introduction

The amount of scientific knowledge that is publicly available has increased dramatically in the past few years. For example, PubMed, a search engine of biomedical publications,[1] now indexes over 25 million papers, 17 million of which were published between 1990 and the present. This information overload yields two critical problems. First, this exceeds the human capacity to aggregate and interpret the fragments of knowledge published in these papers, which may result in existing solutions to critical problems being overlooked. Swanson (1986) described this problem as "undiscovered public knowledge". Second, this vast amount of available information complicates the identification of "**white spaces**" in science, i.e., topics that are insufficiently studied and may lead to important scientific discoveries.

While the first problem has been addressed recently with efforts that combine machine reading and assembly with existing data analysis algorithms (Valenzuela-Escarcega et al., 2017; Poon et al., 2015, inter alia), the second problem is largely unstudied.

In this work we propose a first enabling step towards addressing the problem of white space discovery from literature (Sebastian et al., 2017; Cameron, 2014) with an approach inspired from the field of link prediction (Liben-Nowell and Kleinberg, 2007; Leskovec et al., 2010). In particular, our method operates over two graphs: (a) a graph of positive/negative influence relations such as the relation "CTCF activates FOXA1" between the two proteins, which were extracted using an existing, open-domain machine reading tool (Hahn-Powell et al., 2017) from over 100K biomedical publications[2]; and (b) the citation graph between the corresponding papers where these findings were published. The proposed approach approximates the task of white space discovery by predicting new influence relations that do not exist in the influence graph at a given time (hence the white space) but will emerge in future (thus somebody identified the missing knowledge as important).

The contributions of this work are:

(1) We propose a novel machine learning (ML) framework for this prediction task that uses features extracted from both the influence graph (e.g., the connectivity of relevant concepts in the graph) and the citation graph (e.g., the affinity between related influence relations measured by membership to communities in citation space).

(2) We evaluate the proposed method on an influence graph extracted from over 100K pa-

[1] http://www.ncbi.nlm.nih.gov/pubmed

[2] These relations were extracted using a grammar that identifies causal statements in text. However, we prefer the term "influence" to "causality" in this work because here we simply rely on the text and do not demonstrate that these findings are truly causal.

pers, which contains 1,564,748 concepts (e.g., “astrocytes”, “proinflammatory cytokines”) and 2,395,944 influence relations (e.g., “VEGF increases Akt”). Our method obtains an F1 score of 27 points and a mean average precision of 68%. This outperforms considerably methods that extract features only from one of the two graphs.

(3) To promote future work on this topic, we release a dataset containing both the influence and the citation graphs used in this paper, available at: https://github.com/clulab/releases/tree/master/textgraphs2018-discovery.

2 Data

As mentioned, the primary graph this method operates on is a graph of influence relations extracted from a corpus of 119,667 PubMed Open Access publications. These papers were previously selected to be relevant to the topic of children’s health, which spans multiple domains, and includes issues such as stunting, wasting, and malnutrition.

All these papers were processed using the machine reading and assembly software of Hahn-Powell et al. (2017). In order to address the multi-domain nature of the children’s health, Hahn-Powell et al. followed the OpenIE-style approach of Banko et al. (2007) for entity extraction by considering expanded noun phrases as a coarse approximation of the concepts relevant to the topic. For event extraction, the authors adapted a subset of REACH grammars (Valenzuela-Escarcega et al., 2017) from the biomolecular domain that capture influence statements (e.g., positive and negative regulations). The adaptation removed selectional restrictions on the arguments of each event predicate. That is, they extract any lexicalized variation of “A causes B” where A and B are concepts identified in the entity extraction step. For example, when processing the sentence “Chronic infection may lead to malnutrition and malabsorption”, the system extracts the following entities: “Chronic infection”, “malnutrition”, and “malabsorption”. In this particular case, the extracted entities participate in two `promotes` relations: the first between “Chronic infection” and “malnutrition”, and the second between “Chronic infection” and “malabsorption”.

This machine reading approach was used to read the entire content of these publications (including abstract and body of paper). To reduce noise we kept only relations extracted at least twice, and which occur between concepts with an inverse document frequency (IDF) larger than 1. The resulting influence graph (IG) contains 1,564,748 distinct nodes, connected by 2,395,944 influence relations.

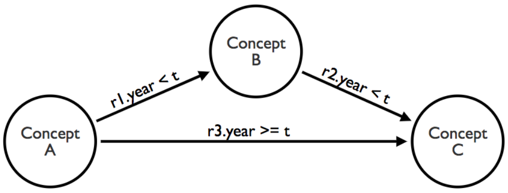

Figure 1: Intuition of the backtesting framework: we predict if links such as A → C will be added to the influence graph after time t, using information that exists before time t such as A → B, and B → C. Setting $t = 2012$ yields 5,015 positive examples, i.e., A → C links that were added after 2012, and 274,251 negative examples, i.e., A → C links that were <u>not</u> added to the graph between 2012 and the end of 2017.

Each match to these rules produces a directed influence relation[3] that encodes polarity (i.e., increase or decrease). Finally, the relation instances are then consolidated through a conservative deduplication procedure.

Because the hypotheses studied in these publications are generally expressed using causal language, we believe this influence graph (IG) captures the essence of the scientific knowledge in this domain.

The above IG is accompanied by a citation graph (CG), which contains outgoing citations from the above papers, at a total of 5,523,759 citation links.

We model the discovery of important white spaces in this knowledge base as link prediction: we predict which influence links will be added to the IG after time t, using only information available before time t. We believe this is a reasonable approximation for the discovery of important white spaces in science knowledge: influence links that will be added in the future indicate that somebody identified the missing information to be important enough to be studied and published. To limit search space, in this work we focus on the prediction of A → C influence links, when A → B and B → C exist in the graph before time t, for at least one node B. Figure 1 visualizes this procedure.

[3] It should be noted that this approach only reads such statements from publications, and does not attempt to verify these findings directly through separate modeling. In other words, the authors statements are assumed to be correct.

t (year)	positive	negative
2017	-	998,586
2016	319	979,709
2015	1,706	881,689
2014	3,767	696,406
2013	5,208	481,671
2012	**5,015**	**274,251**
2011	3,741	151,460
2010	2,448	73,226
2009	1,521	36,839
2008	782	16,372
2007	444	7,843

Table 1: Total number of positive and negative examples for different values of the threshold t.

	Positive Examples	Negative Examples
Training	3,011	164,551
Development	1,002	54,850
Testing	1,002	54,850

Table 2: Number of positive and negative examples in the training, development, and testing partitions.

Note that transitivity cannot be assumed to be true in this graph due to the fact that influence relations extracted from text usually oversimplify complex causal processes. We show in Section 4 that relying on this transitivity assumption leads to poor predictions.

We implemented the above task through back-testing. That is, we look at an arbitrary point in time in the past (t), and create positive training examples from A $\rightarrow$ C links that were added to IG after t. Similarly, we create negative examples from A $\rightarrow$ C links that do not appear between time t and the present. Note that these negative examples are an approximation: some of these may correspond to inventions that will be published at future dates that are beyond the coverage of our dataset. In this paper we used $t = 2012$. This choice of t is justified in Table 1, which shows the distribution of positive and negative examples for different values of t. We chose $t = 2012$ because this provided a large enough number of positive examples, while still maintaining a realistic distribution of negative examples.

The dataset was split into training/development/testing as indicated in Table 2, using a 60-20-20% split. During the partitioning, we made sure that identical influence links coming from different papers are all allocated to the same partition.

3 Approach

We model link prediction as i.i.d. classification on the above dataset, exploring multiple classifiers in Section 4. One key contribution of this work is the feature set used by these classifiers, which is summarized in Table 3. At a higher level, these features capture the connectivity of both the IG and CG around a candidate link, under the assumption that the more connected the corresponding graph is around A and C, the more likely it is that the link $A \rightarrow C$ will be discovered in the near future. In particular, from the IG we extract the in- and out-degrees of the source/destination nodes, and statistics from the path(s) connecting the two nodes such as the length of the shortest path connecting the source and destination nodes, or the inverse document frequency (IDF) scores of the nodes on these paths.

From the CG, we derive features based on the probabilities that papers containing $A \rightarrow B$ ($\mathbf{p}_{A\rightarrow B}$) and $B \rightarrow C$ ($\mathbf{p}_{B\rightarrow C}$) belong to the same community/ies, motivated by the idea that discoveries are easier to be made if the individual fragments that form the puzzle ($A \rightarrow B$ and $B \rightarrow C$ here) come from the same or related discipline(s). We model the probability that two papers, $p1$ and $p2$, belong to the same community $P(p1, p2)$ using two configurations of the Coda community detection algorithm (Yang et al., 2014), one in which detects 100 communities, and another where it detects 300. Because influence links may be reported in more than one paper, we derived the max/min/avg P($p_{A\rightarrow B}$,$p_{B\rightarrow C}$) features, which are computed across all possible combinations of papers $p_{A\rightarrow B}$ and $p_{B\rightarrow C}$.

Lastly, we add a series of features (bottom part of Table 3) extracted from the collection of biomedical publications used in these experiments, such as IDF scores of the relevant concept nodes and the counts for the number of papers that mention a given influence link.

4 Results

Table 4 lists the results of several classifiers on the test partition,[4] compared against two baselines. The first baseline randomly creates positive links following the distribution of positive examples from the training partition. The second base-

[4] All classifier hyper parameters were tuned on the development partition. All classification results were averaged over 5 runs.

Feature Name	Description
C_A.outdegree	Out-degree of source concept node A, i.e., number of influence relations starting on A
C_A.indegree	In-degree of source concept node A, i.e., number of influence relations ending on A
C_C.outdegree	Out-degree of destination concept node C
C_C.indegree	In-degree of destination concept node C
$C_{\text{inbetween}}$.outdegree	Out-degree of nodes in all the shortest paths that connect A to C but do not pass through B
$C_{\text{inbetween}}$.indegree	In-degree of nodes in all the shortest paths that connect A to C but do not pass through B
shortest_path_length	The length of the shortest path that connects A to C but does not pass through B; 0 if no such path exists
shortest_path_count	The number of shortest paths that connect A to C but do not pass through B
$C_{\text{inbetween}}$.avg-idf	Average inverse document frequency (IDF) of nodes in-between A and C in all the above shortest paths
$r_{\text{inbetween}}$.avg-seen	Average number of papers containing an edge in the above shortest paths
max P($p_{A\to B}$,$p_{B\to C}$)	Maximum probability of papers $p_{A\to B}$ and $p_{B\to C}$ being related based on their membership to multi-communities detected by the Coda algorithm; p_r refers to any paper that contains influence relation r.
min P($p_{A\to B}$,$p_{B\to C}$)	Minimum probability of papers $p_{A\to B}$ and $p_{B\to C}$ being related based on their membership to multi-communities detected by the Coda algorithm
avg P($p_{A\to B}$,$p_{B\to C}$)	Average probability of papers $p_{A\to B}$ and $p_{B\to C}$ being related based on their membership to multi-communities detected by the Coda algorithm
Jaccard($\mathbf{p}_{A\to B}$,$\mathbf{p}_{B\to C}$)	Jaccard similarity between the set of papers that contain the link $A \to B$ ($\mathbf{p}_{A\to B}$) and the set of papers that contain $B \to C$ ($\mathbf{p}_{B\to C}$)
Inter-citation ratio	The number of citations between the two sets $\mathbf{p}_{A\to B}$ and $\mathbf{p}_{B\to C}$ normalized by the size of the union of the two sets.
C_A.idf	IDF of the lemmatized terms of source concept node A
C_B.idf	IDF of the lemmatized terms of intermediate concept node B
C_C.idf	IDF of the lemmatized terms of destination concept node C
r1.seen	Number of papers that contain the influence relation A → B
r2.seen	Number of papers that contain the influence relation B → C

Table 3: List of features used by the link prediction classifier that classifies the candidate link $A \to C$, given an intermediate node B. The top part of table lists the features derived from the influence graph. The middle part lists features extracted from the citation graph. The bottom part contains features extracted from the collection of papers.

line assumes that all candidate links are positive, i.e., candidate $A \to C$ is always correct if $A \to B$ and $B \to C$ exist for some B.

	F1	Precision	Recall	P@10	MAP
Baseline (random)	0.02	0.02	0.02	-	-
Baseline (all positive)	0.035	0.018	**1**	-	-
Neural Network	**0.27**	0.398	0.206	0.8	0.537
AdaBoost	0.27	**0.536**	0.178	**0.9**	**0.681**
Random Forest	0.23	0.244	**0.216**	0.5	0.309

Table 4: Unranked scores –precision, recall, and F1– and ranked scores –precision at 10 (P@10), and mean average precision (MAP)– of several classifiers for the prediction of influence links using backtesting at time $t = 2012$. The baseline predicts that every $A \to C$ link will be discovered after time t, if $A \to B$ and $B \to C$ exist before time t.

This table yields several observations. First, the performance of the first (random) baseline is very low, indicating that this is indeed a hard task that is exacerbated by the biased label distribution. Second, the precision of the second baseline is also very low, confirming that the transitive closure assumption is not supported on this realistic influence graph. Third, all classifiers considerably outperform the baseline, indicating that capturing the structure of the IG and CG is indeed indicative of the likelihood that an influence link will be discovered in the near future. Fourth, a linear support vector machines (SVM) classifier did not converge on this data, indicating that, while it is possible to learn a model for this link prediction task, the resulting model is more complex than a linear function. All in all, the best non-linear model (AdaBoost) obtained an F1 score of 0.27, an order of magnitude higher than the baseline, and a mean average precision (MAP) of 0.68, indicating that most correct predictions are ranked closer to the top.

Table 5 shows the results of an ablation experiment in which we measured the drop in performance when each feature was individually removed from the full AdaBoost model. This experiment indicates that, importantly, both the influ-

Removed Feature	F1	Precision	Recall
Full model	0.268	0.528	0.176
$-$ C_A.outdegree	0.234	0.678	0.14
$-$ C_A.indegree	0.246	0.44	0.17
$-$ C_C.outdegree	0.214	0.248	0.186
$-$ C_C.indegree	0.2	0.232	0.172
$-$ $C_{\text{inbetween}}$.outdegree	0.22	0.272	0.182
$-$ $C_{\text{inbetween}}$.indegree	0.234	0.3	0.192
$-$ $C_{\text{inbetween}}$.avg-idf	0.214	0.272	0.18
$-$ $r_{\text{inbetween}}$.avg-seen	0.23	0.29	0.192
$-$ shortest_path_count	0.222	0.29	0.182
$-$ shortest_path_length	0.204	0.28	0.16
$-$ max P($p_{A \rightarrow B}$,$p_{B \rightarrow C}$) (c=100)	0.226	0.3	0.18
$-$ min P($p_{A \rightarrow B}$,$p_{B \rightarrow C}$) (c=100)	0.228	0.302	0.184
$-$ avg P($p_{A \rightarrow B}$,$p_{B \rightarrow C}$) (c=100)	0.232	0.306	0.186
$-$ max P($p_{A \rightarrow B}$,$p_{B \rightarrow C}$) (c=300)	0.232	0.314	0.182
$-$ min P($p_{A \rightarrow B}$,$p_{B \rightarrow C}$) (c=300)	0.23	0.298	0.18
$-$ avg P($p_{A \rightarrow B}$,$p_{B \rightarrow C}$) (c=300)	0.232	0.326	0.182
$-$ Jaccard($\mathbf{p}_{A \rightarrow B}$,$\mathbf{p}_{B \rightarrow C}$)	0.226	0.29	0.184
$-$ Inter-citation ratio	0.23	0.318	0.18
$-$ C_A.idf	0.248	0.478	0.168
$-$ C_B.idf	0.216	0.258	0.19
$-$ C_C.idf	0.22	0.256	0.194
$-$ r1.seen	0.226	0.284	0.19
$-$ r2.seen	0.228	0.298	0.184

Table 5: Ablation experiment, which removed one feature at a time from the full AdaBoost model. This experiment was performed on the development partition.

ence and citation graphs contribute to the overall performance. Removing individual features from either group impacts performance. Several features have a high impact, including $C_{\text{inbetween}}$.avg-idf, shortest_path_length, which are extracted from the influence graph, and max P($p_{A \rightarrow B}$,$p_{B \rightarrow C}$) and Jaccard($\mathbf{p}_{A \rightarrow B}$,$\mathbf{p}_{B \rightarrow C}$), which are extracted from the citation graph. These results demonstrate that the task of scientific discovery requires a multifaceted approach that analyzes several graphs, including graphs that model the content of publications (our IG), as well as citation graphs.

Lastly, we rank the discoveries made by the proposed approach using the NN model, using a scoring function that combines the classifier confidence and redundancy (i.e., how many times we saw $A \rightarrow C$ with different intermediate nodes B) using a Noisy-Or formula:

$$Score(A \rightarrow C) = (1 - \prod_B (1 - prob(A \rightarrow C|B)))$$

where B loops overall intermediate nodes that support the predicted relation, and $prob(A \rightarrow C|B)$ is the classifier's output probability given one intermediate node B.[5] Table 6 lists the top 10 prediction of our approach under this scoring function. 80% of these predictions are correct (i.e., they are discovered after time $t = 2012$). Additionally, the table shows that the predictions are indeed informative: they capture fragments of protein signaling pathways, and links to biological processes (e.g., apoptosis). A few predicted links such as "TLR → cascade" are not informative, but this could be attributed to limitations in the machine reader, which failed to capture meaningful content from the destination concept ("cascade").

Predicted discovery	**Score**	Gold label
antibodies → apoptosis	1	1
apoptosis → ROS	1	1
TGF-beta → apoptosis	1	1
TLR → cascade	1	1
apoptosis → insulin	1	0
apoptosis → enzymes	1	0
antibodies → receptor	1	1
IL-6 → tumor	1	1
mutations → inflammation	0.999	1
macrophages → tumor	0.999	1

Table 6: Top 10 predicted links by the neural network model, sorted in descending order of their informativeness score.

5 Conclusion

We proposed a novel strategy for the identification of white spaces in scientific knowledge, which are topics that are insufficiently studied and may hide important scientific discoveries. We addressed this task with a link prediction method that operates over two graphs: a graph of influence relations that were automatically extracted from over 100K papers on children's health using a machine reading tool, and which summarize the scientific knowledge in this domain, and a graph of citations originating from these papers. Using a backtesting methodology, we showed that our method is capable of predicting which influence links will be discovered in the future with a F1 score of 27 points, and a mean average precision of 68%. An ablation analysis experiment demonstrated that features extracted from both graphs contribute to overall performance. We believe this work is relevant to many actors involved in scientific discovery including researchers and program managers.

Acknowledgements

Marco Valenzuela-Escárcega, Gus Hahn-Powell, and Mihai Surdeanu declare a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies. This work was funded by the Bill and Melinda Gates Foundation HBGDki Initiative.

[5] This ranking function was also used to generate the ranked evaluation in Table 4.

References

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In IJCAI, volume 7, pages 2670–2676.

Delroy Huborn Cameron. 2014. A context-driven subgraph model for literature-based discovery. Ph.D. thesis, Wright State University.

Gus Hahn-Powell, Marco A Valenzuela-Escarcega, and Mihai Surdeanu. 2017. Swanson linking revisited: Accelerating literature-based discovery across domains using a conceptual influence graph. Proceedings of ACL 2017, System Demonstrations, pages 103–108.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Predicting positive and negative links in online social networks. In Proceedings of the 19th international conference on World wide web, pages 641–650. ACM.

David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. journal of the Association for Information Science and Technology, 58(7):1019–1031.

Hoifung Poon, Kristina Toutanova, and Chris Quirk. 2015. Distant supervision for cancer pathway extraction from text.

Yakub Sebastian, Eu-Gene Siew, and Sylvester O Orimaye. 2017. Emerging approaches in literature-based discovery: techniques and performance review. The Knowledge Engineering Review, 32.

Don R Swanson. 1986. Undiscovered public knowledge. The Library Quarterly, 56(2):103–118.

Marco A. Valenzuela-Escarcega, Ozgun Babur, Gus Hahn-Powell, Dane Bell, Thomas Hicks, Enrique Noriega-Atala, Xia Wang, Mihai Surdeanu, Emek Demir, and Clayton T. Morrison. 2017. Large-scale automated reading with Reach discovers new cancer driving mechanisms. In Proceedings of the Sixth BioCreative Challenge Evaluation Workshop, pages 201–203.

Jaewon Yang, Julian McAuley, and Jure Leskovec. 2014. Detecting cohesive and 2-mode communities indirected and undirected networks. In Proceedings of the 7th ACM international conference on Web search and data mining, pages 323–332. ACM.