Topic Influence Graph Based Analysis of Seminal Papers

Abhirut Gupta
IBM Research AI, India
abhirutgupta@in.ibm.com

Sandipan Sikdar
RWTH Aachen University, Germany
sandipan.sikdar@cssh.rwth-aachen.de

Prateeti Mohapatra
IBM Research AI, India
pramoh01@in.ibm.com

Niloy Ganguly
IIT Kharagpur, India
niloy@cse.iitkgp.ernet.in

ABSTRACT

Every scientific article attempts to derive knowledge from existing literature and augment it with new insights. This dynamics of knowledge is commonly explored through references (it draws knowledge from) and citations (its impact on the field). In this paper, we propose to explore this phenomenon through construction of a topic influence graph (TIG) based on topic similarity between articles. More importantly, out of the set of possible TIGs, we determine an optimal TIG by using knowledge from citation graphs. Construction of TIG leverages traditional network analysis tools like community (sub-field) identification. In this paper, we construct the TIG on the ACL Anthology Network (AAN) dataset and leverage it to analyze the properties of seminal papers. Interestingly, we observe that seminal papers disseminate knowledge across different communities, trigger more research within its own community and apart from introducing new ideas, string together ideas from different communities.

CCS CONCEPTS

• **Information systems** → *Document topic models*; *Content analysis and feature selection*;

KEYWORDS

citations, knowledge graph, topic influence graph

ACM Reference Format:
Abhirut Gupta, Sandipan Sikdar, Prateeti Mohapatra, and Niloy Ganguly. 2020. Topic Influence Graph Based Analysis of Seminal Papers. In *7th ACM IKDD CoDS and 25th COMAD (CoDS COMAD 2020), January 5–7, 2020, Hyderabad, India.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3371158.3371191

1 INTRODUCTION

Citation count is the commonly accepted metric for evaluating the impact of a scientific article [2]. Highly cited contributions remain an important criterion for different organizations to identify the best talents. The pattern of citation reflects the nature in which a paper has derived knowledge from previous papers as well as the knowledge it has contributed in the publication of new papers. However, the exact dynamics of knowledge is difficult to comprehend as a citation link is not explicitly connected with any semantics. There are works to annotate the citation links with semantics [4, 9], however in reality, often a cited paper is a representative sample of a set of similar papers; also few relevant (group of) papers are missed out from citations due to various reasons [3] (authors' ignorance, space limitation etc.) - cognizance of these sets would immensely help in understanding the knowledge evolution through scientific articles.

Topic Influence Graph: The latent structure encoding knowledge flow can be represented by a Topic Influence Graph (TIG). We conceive a TIG as one where each paper connects with papers that are 'similar' to its content (topics). To build the TIG, we leverage Latent Dirichlet Allocation (LDA) based topic modeling to obtain the most relevant topics corresponding to each paper. The similarity between two papers is measured as the extent of similarity in topic distributions of the two papers. A similarity threshold is then used to create an edge between the two papers. There is, however, the issue of setting the threshold - we use the existing citation graph to do so. Since a citation graph also connects similar papers, the threshold is set to a value that ensures 'optimal' overlap between the TIG and the citation graph. As we are primarily interested in exploring the potential of TIG towards discerning different aspects of knowledge evolution, we favor a simpler topic similarity mechanism for modeling influence over a more involved probabilistic model [8].

Contrast with existing literature: Extracting relevant topics from scientific articles using topic models has been thoroughly investigated [5, 6], but the research reports have mainly been directed towards understanding evolution of topics (a common use case being given a new topic, from where does it evolve [12]). In fact, citation information has also been leveraged in such endeavors [6, 11]. Further, topic models have been employed for retrieving relevant papers [7]. Although we follow a similar path of extracting topics from papers, our primary contribution is in constructing a topic influence graph based on topic similarity, thereby facilitating the use of elegant network analysis tools in exploring different aspects related to topic evolution.

Dataset: While there are many large datasets with citations and other metadata for scientific publications available in the open domain, most of them do not include the full text of articles. Our approach of building a TIG relies on the topical content of these articles, thus requiring the full text along with its metadata. We use

CoDS COMAD 2020, January 5–7, 2020, Hyderabad, India

ACM ISBN 978-1-4503-7738-6/20/01…$15.00
https://doi.org/10.1145/3371158.3371191

crf sequence conditional fields random crfs label labeling mccallum field sequences
wikipedia relation entity relations extraction entities knowledge articles text web
semantic semantics syntactic meaning representation lexical language representations
learning language acquisition learner learn child input children cognitive model

Figure 1: Top words for the first few topics. At a cursory glance one can see that the topics are about CRF, entity and relation extraction, semantic representation, and learning theory respectively.

the AAN dataset[1], which contains full text of 22,484 papers in the domain of Natural Language Processing from 1965 up to 2014 along with citation, author and venue information. The dataset contains papers from around 18k unique authors spread across 300 venues together accounting for 122k citations.

Seminal Papers: TIG can be utilized to answer several interesting questions on knowledge evolution; in this paper, we take up the issue of characterizing seminal papers. We run a community detection algorithm on the chosen TIG to group papers having similar topics (largely belonging to the same sub-field). On probing the citation and reference patterns of the seminal papers, we observe -

(1) Citations from and references to seminal papers are spread across different communities, while for others they are largely concentrated to one community.
(2) Seminal papers stitch together diverse topics, opening up a new sub-field and thereby encouraging more contributions in this new direction. Hence, seminal papers are often flag-bearers of their corresponding community.
(3) As a proof of concept, we show that the almost all seminal papers are one of the earliest papers to be published in that community.

2 TOPIC INFLUENCE GRAPH CONSTRUCTION

Construction of the TIG is accomplished by a) extracting good topics from papers, and b) constructing edges based on common topics across these papers, which we elaborate next.

2.1 Extract Topics from Papers

We run a LDA based topic modeling on our papers, with 200 topics. Figure 1 shows the top words for the first few topics. We also manually remove a set of stop topics (29 such topics are present in our case) which contain words that don't necessarily reflect an idea or topic from literature, but contain tokens that are frequently found in academic publications. In fact, stop topics are the most frequent ones across documents.

After running LDA, we obtain a representation of each paper by a set of 171 topics with weight (probability) attached to each. Since the probability of most of the topics would be quite low, we would like to determine the number of topics which would be enough to faithfully represent a paper. To investigate the contribution of each topic in determining the property of a node (paper), we plot the cumulative probability distribution versus the number of topics (averaged over all papers) in Figure 2. We observe that the top five

[1]http://tangra.cs.yale.edu/newaan/

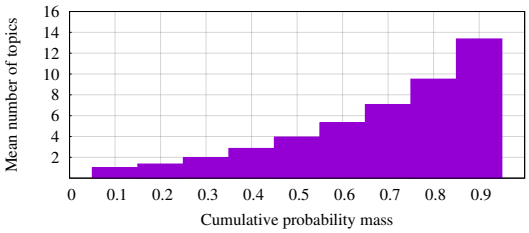


Figure 2: Average Cumulative Probability across papers against number of topics

topics contribute ~60% of the probability mass - a topic beyond top five contribute only minimally. Hence, based on their impact, we posit that a paper can be meaningfully represented by the top five topics.

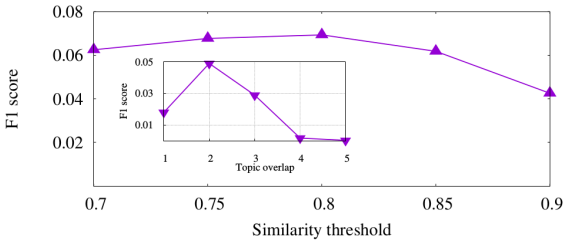
2.2 Constructing Edges

In order to construct a TIG, a similarity measure needs to be defined between two nodes - we formulate two similarity measures. The first is **soft similarity**, where we draw an edge between two papers based on the thresholded similarity of their topic distributions. That is, since each paper is represented by a probability distribution of the constituent topics, we draw an edge between two nodes if their *Spearman correlation coefficient* is above a threshold. The other one is **hard similarity**, where we connect two papers if they have a minimum number of topics in common. In general, the graphs obtained through the first measure are more conducive to our analysis (details provided in the next section).

2.3 Determining the Optimal Graph

As we would obtain a set of graphs by varying the threshold and the similarity measure, we need to chose the 'best' graph from the set. Citation graph can be conceived as a human-sampled topic influence graph and can act as a guiding structure in determining the quality of a TIG. We hence postulate that the selected graph should have the 'best' overlap with the underlying citation graph. We calculate precision and recall of the citation graph in a TIG as Precision = $\frac{n}{k}$, Recall = $\frac{n}{c}$, where n is the number of citation edges covered by TIG, c is the total number of edges in the citation graph, and k is the total number of edges in the TIG. Subsequently, the value of F_1 can be calculated - we posit the overlap between the citation graph and a TIG is best determined by the F_1 score as it balances precision and recall.

To obtain the optimal overlap between the topic influence and the citation graph, we calculate the F_1 score of the overlap for different threshold values (Figure 3 (top)) and observe that the highest score is obtained for a threshold value of 0.80 for similarity beyond which the graph becomes too restricted. The value for 0.75 (SSim_s0.75) is almost same as 0.80. However, (SSim_s0.75) is a better connected graph (Figure 3 (bottom)) including around 90% of the nodes - hence we shortlist this graph for our future analysis. We also experimented with different values in hard similarity, but the F_1 scores were always less compared to the so similarity cases (Figure 3 **top** (inset)). All further results in the paper are presented on the graph SSim_s0.75 (referred to as TIG_O) with the assumption that each node is represented by a set of five topics (determined in Section 2.1).



Graph Name	Number of Nodes with At least One Edge	Number of Edges	Precision	Recall	F_1
SSim_s0.80	17997	370877	0.046	0.139	**0.069**
SSim_s0.75	20583	605743	0.041	0.201	0.068
SSim_s0.70	21872	921144	0.035	**0.266**	0.063
SSim_s0.85	14106	203463	0.049	0.082	0.062
SSim_s0.90	8972	91504	**0.049**	0.037	0.042
Citation Graph	-	122403	-	-	-

Figure 3: (Top)F1 scores for So Similarity graphs with varying similarity thresholds. (inset) Same result for Hard Similarity with X-axis representing number of topics in common. General properties related to the graphs obtained with different threshold values are summarized in the table (below)

Table 1: Top five papers in our sample. Note that the scores are normalized between 0 and 1 and higher the score for a paper, higher are its chances of being seminal.

Title	Year	Score	Citations
Moses: Open Source Toolkit for Statistical Machine Translation	2007	1.0	737
Building A Large Annotated Corpus Of English: The Penn Treebank	1993	0.99	989
Bleu: A Method For Automatic Evaluation Of Machine Translation	2002	0.983	1055
A Systematic Comparison Of Various Statistical Alignment Models	2003	0.976	746
Attention, Intentions, And e Structure Of Discourse	1986	0.97	369

3 SEMINAL PAPER ANALYSIS

In this section, we use TIG_O to investigate the properties of seminal papers and understand the unique properties manifested by these papers.

3.1 Selecting the Seminal Papers

We use the work proposed by [10] for selecting the seminal papers. They observe that the citations have a fat-tailed distribution, with papers with long-term impact accounting for the fat tails. For each paper, we calculate a "seminality" score considering the average fraction of citations it receives over the years after its publication. The rationale for averaging is to eliminate the additive effects of citation arising from the years that have passed after the publication of the paper. The average score unbiasedly reflects the importance of a paper, discounting the aging effect of a paper. We select the top 100 papers based on the seminality score for our analysis (Table 1 gives the top 5 papers). Note that higher citations does not always lead to higher seminality as evident in table 1.

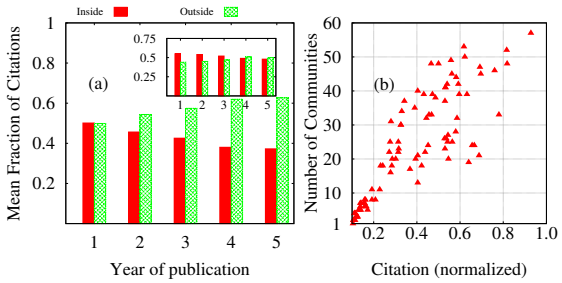


Figure 4: (a) Fraction of citations from inside and outside the community from the year of publication. X-axis represents the year from the publication and Y-axis represents the fraction of citations averaged across all the seminal papers, (inset) sample plot with non-seminal papers. (b) Citation count versus number of communities citation is obtained from for all the seminal papers. Higher the count higher is the number of communities a seminal paper influences.

3.2 Determining communities

We conceive that the seminal papers would have a strong impact on sub-fields, may start a new sub-field etc. We posit that the communities imbibed inside TIG_O roughly represent subfields. We run the Louvain algorithm [1] on TIG_O to detect communities. Louvain returns 217 communities with each paper being member of a unique community, and a modularity score of 0.887. The largest community has 1716 papers, and the smallest has 2. Manual inspection reveals that largely each community contains paper predominantly of a subfield (rigorous experiment not done). Topic clouds related to the papers in two sample communities are presented in Figure 5. Note that they represent the subfields of machine translation and word-sense disambiguation respectively.

3.3 Properties of Seminal Papers

Based on this community information, we intend to understand how seminal papers draw and disseminate knowledge compared to non-seminal papers.

Property 1 - Disseminating knowledge across communities: The first property we study is the pattern of citations of seminal papers across communities. We find that on an average seminal papers are cited across 26 communities, compared to only 3 for non-seminal papers. This is, however, expected as seminal papers are in general more cited than other papers. The more interesting result is the pattern of inside and outside community citations. Figure 4 (a) shows the fraction of citations from inside and outside that community from the year of publication averaged across all seminal papers. Interestingly, we observe that the fraction of citations from outside the community increases over time indicating its long term influence. We also observe a direct relation between seminality of a paper and the number of communities citing the paper. This is inferred from Figure 4 (b) which presents the number of citations a paper accrued against the number of communities it received

Figure 5: Word cloud for topic words from two sample communities - the first one has papers predominantly on machine translation and the second one has papers from topic of word sense disambiguation.

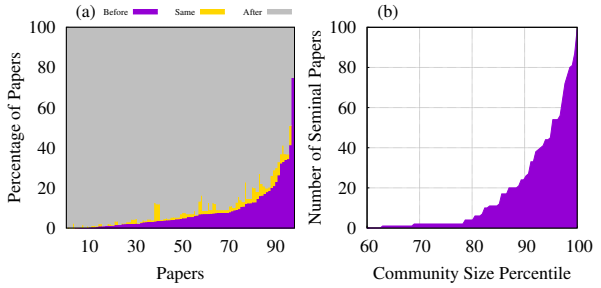


Figure 6: (a) Percentage of papers before (purple), in the same year (yellow), and after seminal papers (grey) in their communities. (b) Cumulative Distribution of seminal papers vis-a-vis community size.

citations from (i.e., the number of communities it could influence) for all the seminal papers.

Property 2 - Temporal position in the community: We analyze the temporal positioning of seminal papers in their respective communities to understand if seminal papers drive a community. Figure 6 (a) shows the percentage of papers in the community published before, in the same year, and after a seminal paper's year of publication. We can clearly see that an overwhelming majority of the seminal papers are published before the bulk of the papers in that community appear. Similarly for non-seminal papers we observe a major bulk of the papers (45.5% on average) in the community are published before. Also we check the size of the communities to gauge the extent to which the papers have initiated the scientific activities. Figure 6 (b) reflects that the distribution of seminal papers is heavily skewed towards the larger communities. This shows that seminal papers start and sustain larger and important sub-fields. Examples of two seminal papers which are at the start of communities in Figure 5 are "Discriminative Training And Maximum Entropy Models For Statistical Machine Translation (2002)" for the statistical machine translation community and "Word-Sense Disambiguation (WSD) Using Statistical Models Of Roget's Categories Trained On Large Corpora (1992)" for the WSD community.

Property 3 - Stitching together ideas from different fields: As observed before, a seminal paper is often a flag bearer of the community introducing new ideas and thereby determining the direction of future research in the field. We further argue that a seminal paper might also string together ideas from different fields triggering further research. As a proof of concept, we consider a pair of topics (from 5 topics) associated with each seminal paper, published in the year t, and get the number of papers (having the same pair of topics) which were published before and after t. The pair is chosen based on their prevalence in the entire dataset. We observe that across all the seminal papers and pairs of topics, the mean number of papers (having the same pair of topics) published before t is significantly ($p < 0.01$) less compared to the mean number of papers published after. Note this is beyond the necessary course correction taken to ofset the trend of increasing number of overall publications over years. Figure 7 plots the distribution of the (normalized) number of papers over time (i.e., published in ..., t - 2, t - 1, t + 1, t + 2, ...) for a pair of topics for both seminal and non-seminal papers. For seminal papers, we observe an increasing trend with the count spiking just after the publication of the paper (the results are averaged over all the seminal papers). For non-seminal papers however, we see no such spike, indicating that most non-seminal papers are published when a trend already exists. The above observations hence indicate that the seminal papers are indeed able to meaningfully combine ideas from diverse fields, and are early in this effort.

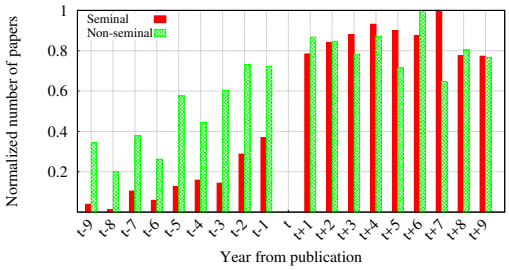


Figure 7: The number of papers (with a given pair of topics) published in years $t-9, ..., t-1, t+1, ..., t+9$ with the paper in focus published in t^{th} year, for both seminal and non-seminal papers.

4 DISCUSSION

In this paper, we propose a novel method of constructing a topic influence graph - the novelty lies in leveraging the citation graph to create the optimal version. The TIG has been used to analyze the properties of seminal papers. Our proposal leads us to identify three factors which might have led to a paper becoming seminal - (i) disseminating knowledge across communities, (ii) triggering more research in the community and (iii) introducing new ideas or meaningfully combining ideas from different areas. We believe that the topic influence graph could lead to the use of enormous network analysis literature in several diverse retrieval tasks. Such investigations call for additional research efforts which we intend to take up in future. In future work, we plan to use Article Influence Score, SCImago Journal and Country Rank in addition to the citation graph for analysis of the seminal nature of papers.

REFERENCES

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[2] Tanmoy Chakraborty, Suhansanu Kumar, Pawan Goyal, Niloy Ganguly, and Animesh Mukherjee. 2014. Towards a stratified learning approach to predict future citation counts. In *JCDL*. IEEE Press, 351–360.

[3] Valerio Ciotti, Moreno Bonaventura, Vincenzo Nicosia, Pietro Panzarasa, and Vito Latora. 2016. Homophily and missing links in citation networks. *EPJ Data*

Science 5, 1 (2016), 7.

[4] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware Citation Recommendation. In *WWW*. ACM, New York, NY, USA, 421–430. https://doi.org/10.1145/1772690.1772734

[5] Yookyung Jo, John E Hopcroft, and Carl Lagoze. 2011. The web of topics: discovering the topology of topic evolution in a corpus. In *WWW*. ACM, 257–266.

[6] Yookyung Jo, Carl Lagoze, and C Lee Giles. 2007. Detecting research topics via the correlation between graphs and texts. In *SIGKDD*. ACM, 370–379.

[7] Ziyu Lu, Nikos Mamoulis, and David W Cheung. 2014. A collective topic model for milestone paper discovery. In *SIGIR*. ACM, 1019–1022.

[8] Benyah Shaparenko and Thorsten Joachims. 2007. Information genealogy: uncovering the flow of ideas in non-hyperlinked document databases. In *SIGKDD*. ACM, 619–628.

[9] Jie Tang and Jing Zhang. 2009. A Discriminative Approach to Topic-Based Citation Recommendation. In *PAKDD*. Springer-Verlag, Berlin, Heidelberg, 572–579. https://doi.org/10.1007/978-3-642-01307-2_55

[10] Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying long-term scientific impact. *Science* 342, 6154 (2013), 127–132.

[11] Xiaolong Wang, Chengxiang Zhai, and Dan Roth. 2013. Understanding evolution of research themes: a probabilistic generative model for citations. In *SIGKDD*. ACM, 1115–1123.

[12] Ding Zhou, Xiang Ji, Hongyuan Zha, and C Lee Giles. 2006. Topic evolution and social interactions: how authors effect research. In *ACM CIKM*. ACM, 248–257.