The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
Scholarly Big Data: AI Perspectives, Challenges, and Ideas:
Technical Report WS-16-13

Encoding Lineage in Scholarly Articles

Sheikh Motahar Naim[1], Md Abdul Kader[1], Arnold P. Boedihardjo[2], M. Shahriar Hossain[1]
[1]Department of Computer Science, University of Texas, El Paso, TX 79968
[2]U. S. Army Corps of Engineers, Alexandria, VA 22315
{snaim, mkader}@miners.utep.edu, arnold.p.boedihardjo@usace.army.mil, mhosssain@utep.edu

Abstract

The development of new scientific concepts today is an outcome of the accumulated knowledge built over time. Every scientific domain requires understanding of the trends of the dependencies between its subdomains. Analyses of trends to capture such dependencies using conventional document modeling techniques is a challenging task due to two reasons: (1) conventional vector-space modeling based representation of documents does not realize the history of the content, and (2) neither feature-level nor document-level causality is provided with any digital library metadata or citation network. In this paper, we propose an intuitive temporal representation of a scientific article that encodes inherent historic characteristics of the content. This intuitive representation of each document is then leveraged to discover causal relationships between scientific articles. In addition, we provide a mechanism to explore the lineage of each document in terms of other previously published documents, which illustrates how the theme of the document under analysis evolved over time. Empirical studies reported in the paper show that the proposed technique identifies meaningful causal relationships and discovers meaningful lineage in the scientific literature that could not be discovered through the citation network of the articles.

1 Introduction

With the rapid growth in electronic publication of scientific articles, we now have many digital libraries richer than ever. For example, both IEEE Xplore (ieeexplore.ieee.org) and Pubmed Central (ncbi.nlm.nih.gov/pmc) have over 3.6 million full-text articles in their collection while ACM Digital Library (dl.acm.org) contains more than 400 thousands publications in computing and information domain. Due to such vast collection of searchable scientific articles, researchers in any domain find it much easier to retrieve documents related to a particular topic in the literature. Almost all of the available digital libraries return search results based on the textual similarity of the initial query. However, results returned against a search query are not enough to realize how the topic evolved and conceptually diffused over time from another topic. The topics of other domains that caused or influenced the state in the past may not contain the same keywords that were used for the query. In this paper, we describe a mechanism that goes beyond similarity to capture the actual dependencies between scientific articles to understand how a document published recently has reached its state.

There are a few attempts to capture the evolution of the current state both at document (Shaparenko and Joachims 2007; El-Arini and Guestrin 2011; Hasan et al. 2009) and concept levels (Blei and Lafferty 2006; Mei and Zhai 2005; Wang and McCallum 2006). Shaparenko and Joachims (Shaparenko and Joachims 2007) explain a mechanism named *information genealogy* that heavily depends on textual similarity to compute dependence between documents while forming a lineage. As a result, the approach does not capture influential documents that do not contain much textual overlap with the initial set of documents. In practice, two scientific articles may contain different textual content but one may influence or relate to the other historically. For example, with the use of terms related to bipolar junctions in physics the vocabulary in electrical engineering started to change which eventually resulted in a strong branch of scholarly endeavor — computer science. Similarly, the innovations as well as the vocabulary in laser physics fostered the area of cancer treatment influencing the literature in biomedical science. The approach we describe in this paper encodes historical trends of the entities used in a paper in the form of a temporal series that is used to detect causality with other articles avoiding direct similarity computation between documents while discovering a lineage. There are a few systems that have the ability to provide an illustration of the conceptual evolution of the literature, such as, Dynamic Topic Modeling (Blei and Lafferty 2006), Online LDA (AlSumait, Barbará, and Domeniconi 2008), and Mei and Zhai's evolutionary theme pattern (Mei and Zhai 2005). While these approaches focus on lineage at conceptual level for an entire corpus, our center of attention in this paper is at document level lineage construction based on the latent causality between the documents of a corpus.

Our framework, through a number of experiments conducted over 400,000 publication abstracts from IEEE Xplore digital library, discovers lineage of documents that similarity-dependent methods cannot detect. The main contributions of this paper are as follows:

1. We represent a document as a time series that encodes historic importance of the terms of the document from the entire vocabulary perspective.

2. Our framework provides a mechanism to cluster documents with high causal relationship.
3. We propose a systematic way to track the lineage of any published article in the form of a chain of causal documents.

2 Related Work

Big collections of scholarly articles from different digital libraries have been exploited in multitude of applications including keyword extraction (Caragea et al. 2014), citation recommendation (Kataria, Mitra, and Bhatia 2010) and summarization of new contributions (Teufel and Moens 2002). In an attempt to find the influential articles of a current document, Shaparenko and Joachims (Shaparenko and Joachims 2007) try to explain the content of a document using the textual content of previously published articles. Though this method performs well to identify influential articles in the literature, it is heavily dependent on textual similarity of the documents, and therefore, fail to identify true influence where there is less or no textual overlap. El-Arini and Guestrin (El-Arini and Guestrin 2011) and Hasan et. al. (Hasan et al. 2009) go beyond the keyword matching and utilize other meta-information like citations to find the related articles for a given set of papers. Another path of work towards understanding the evolution of research has been through incorporating temporal information into topic modeling. Mei and Zhai (Mei and Zhai 2005) find the important themes in every time frame of the corpus and connect those themes based on their thematic similarity to show the evolutionary transitions of different topics. Dynamic Topic Model (Blei and Lafferty 2006) and Online LDA (AlSumait, Barbará, and Domeniconi 2008) also track the topics over time and provide a means to identifying emerging topics in various time points. In this paper we introduce the concept of causality to understanding the evolutionary nature of research. Though the idea of causality has been successfully used in the fields of economics (Granger 1969; Cheng and Lai 1997) and neuroscience (Roebroeck, Formisano, and Goebel 2005), this is the first attempt, to the best of the authors' knowledge, to exploit Granger causality (Granger 1969) for understanding the lineage of scientific documents.

3 Problem Formulation

Let $D = \{d_1, d_2, \ldots, d_N\}$ be the scholarly dataset of N articles containing M words $W = \{w_1, w_2, \ldots, w_M\}$. Each article may contain an arbitrary number of words in any sequence. The publication dates of the documents span over a time frame $Y = \{y_1, y_2, \ldots, y_l\}$ where y_i is ith year and $y_{i+1} = y_i + 1$. We denote the set of articles published in and before year y_i as D_i. The tasks are:

- Identify the set of all causal dependencies in the corpus, $\mathcal{R} = \{(d_i, d_j) : d_j$ has a causal dependence on $d_i\}$.
- Construct a causal chain $H_i = \{h_i^1, h_i^2, \ldots, h_i^{|H_i|}\}$ for each article $d_i \in D$ where $h_i^1 = \{d_i\}$ and h_i^q is the set of documents each of which has causal influence on at least one of the documents in h_i^{q-1}.

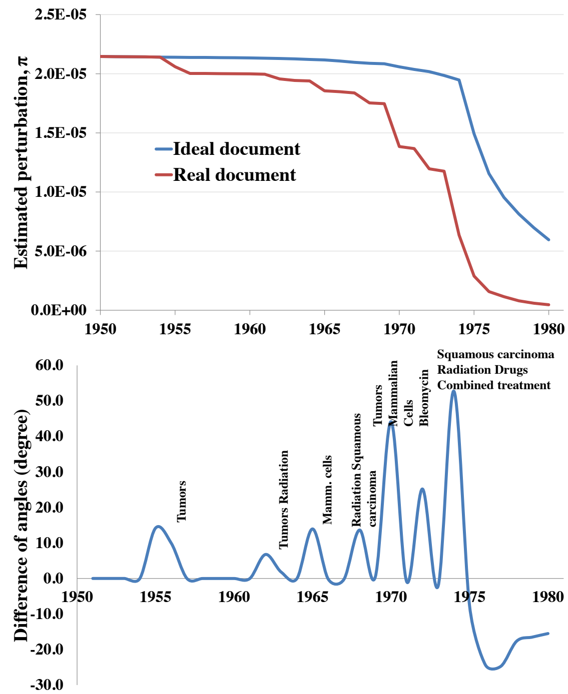

Figure 1: Time series generation for a document. (a) Perturbation caused by a real and the corresponding ideal document. (b) Generated signal for the real document.

- Compute a clustering, $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{|\mathcal{C}|}\}$ of the corpus where documents in each group $\mathcal{C}_i \in \mathcal{C}$ demonstrate strong causal dependence among themselves.

4 Methodology

The objective of the proposed framework is to identify causal dependency structure of all the publications. The following subsections describe the functionalities of the proposed framework: 1) represent documents as time series, 2) identify causal groups of articles, and 3) discover causal chains.

4.1 Document Representation as a Time Series

Our framework captures the change of the distribution of the vocabulary W over time by reading the articles in a sequence as they were published. We compute the term distribution at the end of every year to capture this evolving nature of the vocabulary. Let the term distribution after the ith year be $\gamma_i = \{\gamma_i^1, \gamma_i^2, \ldots, \gamma_i^M\}$ where γ_i^j is the frequency of the jth term of the vocabulary W in all the document published in or before the ith year and $\sum_{j=1}^{M} \gamma_i^j = 1$.

This evolving distribution of the vocabulary enables us to generate a time series for each document based on the relative novelty of that article. If an article $d \in D$ published in the jth year contains entities that have already been published in the literature in any ith year where $i < j$, then d will not introduce much change in γ_i since the distribution already contains some values for those entities. Based on this concept, we replicate document d and place it in the partial corpus D_i. Let the extended corpus for document d in the ith year be $\bar{D}_i = D_i \cup d$ and the term distribution of the extended corpus $\bar{D}_i$ be $\bar{\gamma}_i$. The amount of perturbation that document d creates in the distribution of the ith year is estimated as the Kullback-Leibler divergence between the original distribution γ_i and the distribution of the extended corpus $\bar{\gamma}_i$.

$$\pi_d^i = \sum_{i=1}^{l} \gamma_i \ln \frac{\gamma_i}{\bar{\gamma}_i} \tag{1}$$

The order of two distributions γ_i and $\bar{\gamma}_i$ in Equation 1 is important due to the asymmetric nature of KL-divergence. Since our intention is to measure the extra information added to the corpus D_i by the inclusion of document d, we always compute the KL-divergence from γ_i to $\bar{\gamma}_i$, not the other way around.

By placing d in each year $y_i \in Y$ we get a time series $\pi_d = \{\pi_d^1, \pi_d^2, \ldots, \pi_d^l\}$ for d that represents the document's influence on the vocabulary over all the years. The red line in Figure 1(a) shows this perturbation-based time series of a sample document.

In the next step, we create a synthetic document d' that is an identical copy of d but each of the terms being encoded by a unique identifier. That is, the terms in the synthetic document do not appear in any of the documents in the entire corpus D. This ensures that d' independently holds the properties of d as a single document but it is a novelized one since none of the documents contains the synthetic terms. Now following the same procedure as we used for d, we add d' to D_i to form an extended corpus $\bar{D}'_i$, calculate the term distribution $\bar{\gamma}'_i$ for that corpus, and estimate the amount of perturbation d' could introduce. We construct $\pi_{d'} = \{\pi_{d'}^1, \pi_{d'}^2, \ldots, \pi_{d'}^l\}$ for d' the same way as we computed π_d. The blue line in Figure 1(a) is the signal generated by placing d' in every year. Since the content of d' is unique, the line for d' will always have larger (or equal) value in every year than the line for the real document d.

The time series $\pi_{d'}$ is a representation of the degree of influence the document d would have had on the vocabulary had it be a completely unique document. Therefore, the difference between the two time series π_d and $\pi_{d'}$ gives us an estimation of the amount of uniqueness of document d in every year of the corpus — the smaller the difference the closer it is to the extremely unique document. Figure 1(b) shows the signal for d that is derived from the difference between π_d and $\pi_{d'}$. We denote this signal as S_d and calculate its magnitude in ith year as the difference between the angles created by the time series π_d and π'_d at the ith year.

4.2 Identification of Causal Clusters

The information provided by an article published today is an outcome of knowledge accumulated over time in the past. Each article in a corpus of a certain domain is very likely to have a few articles to which it is causally dependent. We use the signal S_d generated for each document $d \in D$ to compute the causalities between all articles. We leverage Granger causality test in this purpose. Granger causal-

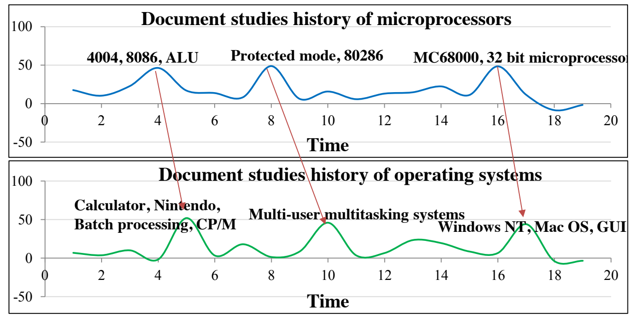

Figure 2: An illustrative example of causality between two documents based on their signals.

ity (Granger 1969) is a statistical hypothesis test which estimates the usefulness of one time series in predicting the future values of another time series. A time series X is said to Granger-cause another time series Y if it can be shown that those X values provide statistically significant information about future values of Y. We use Granger causality to test if there exists any causal relationship between two documents. Assuming that the document d_i was published before the document d_j, we test the following hypothesis to identify a causal effect from d_i to d_j:

$$\mathbb{P}[d_j(t+1) \in A | \mathcal{I}(t)] \neq \mathbb{P}[d_j(t+1) \in A | \mathcal{I}_{-d_i}(t)] \tag{2}$$

where $\mathbb{P}$ indicates probability, A is an arbitrary non-empty set, and $\mathcal{I}(t)$ and $\mathcal{I}_{-d_i}(t)$ respectively denote the information available as of time t in the corpus, and that in the corpus excluding d_i. If the above hypothesis is accepted, we say that document d_i Granger causes document d_j. Figure 2 demonstrates the causal relationship between two documents based on their time series.

Once we identify the set of all the causal relationships $\mathcal{R}$ in the corpus, we can build a causality network or a causality matrix. In practice, for large corpus we limit causality computation for any document to the documents published in r previous years only. The set of IEEE publications that we use in the experimental results section spans over 54 years, and we vary the value of that look-back threshold r from 8 to 10 years for our experiments.

The strengths of the causal relationships in $\mathcal{R}$ allow us to use clustering algorithms to find causal groups of articles. We use a density based clustering algorithm DBSCAN to group causal documents. The motivation behind the use of a density based clustering is that this specific family of clustering algorithms does not require the number of clusters k as an input (unlike k-means clustering). DBSCAN is a logical choice for causal partitioning since there is no metadata that can help us identify possible number of causal clusters.

4.3 Computing Causal Chains

Algorithm 1 provides a chain of causal documents for a given document d_i. Each chain is a graph or tree containing causality flows from leaf documents toward the document for which the lineage is generated. The chain is initialized in line 1-2 with the given document d_i. We expand the chain

Algorithm 1: *ComputeLineage* – algorithm to compute the causal chain of a document.

input : Document d_i
parameter: Set of causal pairs, $\mathcal{R}$
Look-back threshold, r
Branching factor, b
output : Lineage of d_i, h_i

create an empty list h_i;
append $\{d_i\}$ to h_i;
do
  create an empty set S_p;
  $S \leftarrow$ read the last set from h_i;
  for *each document* $d \in S$ **do**
    $\mathcal{R}_d \leftarrow \{(d_p, d_q) : (d_p, d_q) \in \mathcal{R}_d$ and $d_q = d\}$
    if $|\mathcal{R}_d| > b$ **then**
      $\mathcal{R}_d \leftarrow$ top b pairs of $\mathcal{R}_d$;
    end
    for *each causal pair* $(d_p, d_q) \in \mathcal{R}_d$ **do**
      if *yearOf(d_q) - yearOf(d_p)* $\leq r$ **then**
        add d_p to S_p;
      end
    end
  end
  if $|S_p| \neq 0$ **then**
    append S_p to h_i;
  end
while $|S_p| \neq 0$;
return h_i

by adding new set of documents that have causal influence on the documents of the previous set. Each document d in a level is expanded to b articles with highest causalities on d which were published within r years of the publication date of d. The procedure terminates when there is no more causal parent left to expand from a certain level.

5 Evaluation

We evaluate our framework by evaluating the quality of causal clusters, how much information is diffused over time in a lineage, and comparing the lineage produced by our approach with a similarity based model, information genealogy (Shaparenko and Joachims 2007), and a citation network based lineage.

One of the evaluation metrics we use is diffusion coefficient, which estimates a quantity of how the theme changed over time. The basic assumption here is that a lineage of scholarly articles should diffuse a concept over the years and generate new ideas. Let $P = \{p_1, p_2, \ldots, p_{|P|}\}$ be the set of paths in the lineage of a document $d \in D$, and $p_k = \{d_0^{p_k}, d_1^{p_k}, \ldots, d_{n-1}^{p_k}\}$ be the kth path in P containing n documents and $d_0^{p_k}$ refers to the source document d. Diffusion coefficient $\mathcal{D}(d)$ of lineage of the document d is then defined as

$$\mathcal{D}(d) = \frac{1}{|P|} \sum_{k=1}^{|P|} \left(1 - \frac{1}{n-2} \sum_{i=0}^{n-3} \sum_{j=i+2}^{n-1} disp(d_i^{p_k}, d_j^{p_k}) \right) \tag{3}$$

where

$$disp(d_i^{p_k}, d_j^{p_k}) = \begin{cases} \frac{1}{n+i-j} & d_i^{p_k} \text{ and } d_j^{p_k} \text{ have term overlap} \\ 0 & \text{otherwise.} \end{cases}$$

Larger values of $\mathcal{D}(d)$ indicate better diffusion over time while smaller values will refer to lesser conceptual drift.

In addition, we examine how much of the lineage generated by our approach overlaps with the citation tree formed for a document. Let $T^{citation}(d)$ and $T^{causality}(d)$ are the ancestor trees of depth L of a document $d \in D$ generated using the actual citations and causal relationships, respectively. Then a citation overlap score for article d can be defined as:

$$\mathcal{A}^{causality}(d) = \sum_{l=1}^{L} \left| \eta_l^{T^{citation}(d)} \cap \hat{\eta}_l^{T^{causality}(d)} \right| \times \frac{l}{L} \tag{4}$$

where $\eta_l^{T^{citation}(d)}$ and $\hat{\eta}_l^{T^{causality}(d)}$ are the set of references of documents at level l of the tree $T^{citation}(d)$ and $T^{causality}(d)$ respectively. The same equation can be used to compute citation overlap score $\mathcal{A}^{similarity}$ for a similarity based lineage as well as $\mathcal{A}^{genealogy}$ for an information genealogy based approach.

We express the degree of causality within the documents in the same topical group t as causal density score $\rho(t)$:

$$\rho(t) = \frac{2 \times |\mathcal{R}_t|}{n_t \times b} \tag{5}$$

where b is the maximum number of causal parents considered for each document, $\mathcal{R}_t \subset \mathcal{R}$ is the set of causal relations within topic t and n_t is the number of documents in t. We also calculate the causal influence of a topic t_1 on another topic t_2 as:

$$\phi(t_1, t_2) = \frac{|\{(d_i, d_j) \in \mathcal{R} : d_i \in D^{t_1}, d_j \in D^{t_2}\}|}{n_{t_1} + n_{t_1}} \tag{6}$$

where D^{t_1} and D^{t_2} are the set of documents in topic t_1 and t_2 respectively. Finally, the causal dominance $\psi(t)$ of a topic over the other topics in the corpus is calculated as

$$\psi(t) = \frac{1}{n_t} \sum_{d \in D^t} \sum_{t' \in T} \sum_{d' \in D^{t'}} \begin{cases} 1 & (d, d') \in \mathcal{R} \\ 0 & otherwise. \end{cases} \tag{7}$$

where T is the set of all topics

6 Experimental Results

In this section, we seek to answer the following questions to justify the capabilities and correctness of the proposed model.

1. How does the causality based clustering mechanism compare to a baseline similarity based clustering approach? (Section 6.1)

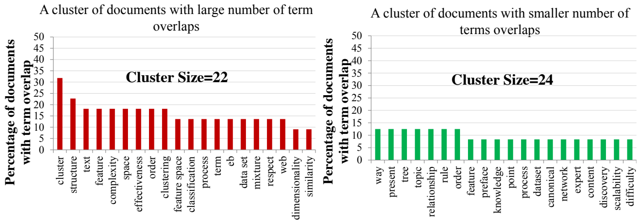

Figure 3: Comparison of two highly causal clusters in term of the textual similarities among the documents of each cluster.

2. Can the proposed lineage formation approach identify genealogies drifted from other topics? (Section 6.2)
3. Which topical groups have most causal influence on the other topical groups in the entire corpus? (Section 6.3)

We collected a publication dataset of titles and abstracts along with some meta-data that includes publication year and citations of each paper. The collection contains 412,484 computer science articles from the IEEE Xplore digital library. The computer science articles are recognized by entries available in the DBLP computer science bibliography database. This collection contains documents from the year 1961 to 2014. We extracted over eight hundred thousand entities from the titles and the abstracts using lingpipe, Stanford NER and openNLP entity detectors (Hossain et al. 2012). The entities were then tokenized to construct the feature set for each document.

6.1 Causality based vs. Similarity based clustering

One may argue that one particular scholarly article is only motivated by similar articles published in the past. While similarity is a good way to discover articles published on the same topic, the inter-topic influence cannot be captured using similarity search. To verify whether highly causal document clusters are always similar or not, we analyze the similarity between the documents of each of the causal clusters detected by density based clustering (as described in Section 4.2). Figure 3 shows the percentage of documents that share certain terms in two different causal clusters. The terms in

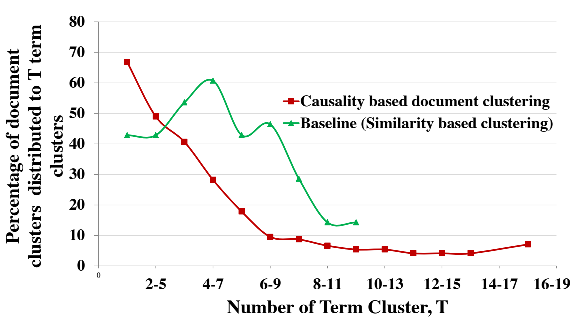

Figure 4: Comparison between causality based and similarity based document clustering with respect to the causality based term clustering.

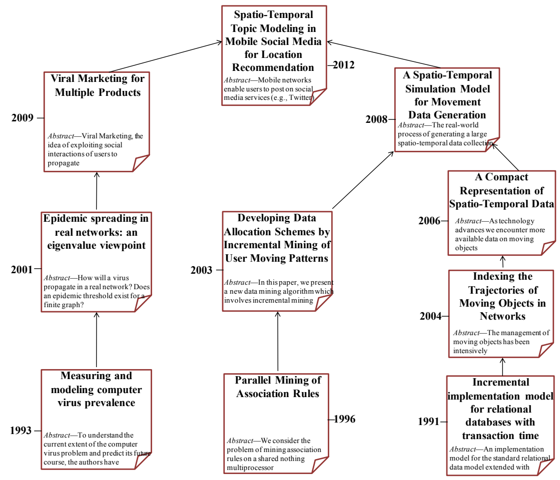

Figure 5: The causal chain of a document discovered by the proposed method.

the x-axis are ordered based on percentage of documents in a cluster containing those terms. The top twenty frequent terms are shown for each cluster. The plot in the left shows that the documents in a highly causal cluster have large amount of term overlaps. For example, 32% and 23% of the documents contain the terms *cluster* and *structure*, respectively. In contrast, the plot at the right side illustrates that the documents in a highly causal cluster can have comparatively low textual similarity. The most frequent term "way" appears in just 12.5% of the documents of the causal cluster represented by the plot at the right side of Figure 3. This stipulates that a causal cluster of documents may or may not be similar in terms of textual contents.

While Figure 3 shows an evidence that similarity is not the key indicator of causality, it does not confirm that highly causal documents grouped in a cluster are causal at the feature level. To examine whether a causality based grouping of documents brings more causal items together than a similarity based grouping, we cluster all the documents based on causality and similarity separately. Then we cluster all the terms of the corpus based on term causality considering the frequency of each term in each year as the amplitude of the corresponding term signal. Additionally, we apply topic modeling to the corpus and assign a topic to each of the articles. We seek to verify the causality of the terms of the documents of a causality based cluster by examining the dominant terms of the topics of the documents of that causal cluster. Those dominating terms should come from a small number of causality based term clusters. In contrast, the dominant topical terms of a similarity based cluster of articles will tend to come from multiple causal term clusters if causality and similarity are less relational. Figure 4 shows that the dominant topical terms of causality based document clusters are distributed to small number of causal term clusters indicating that causality based document clusters are highly causal at term level. The similarity based document clusters exhibit a different trend. The topical terms of the similarity based article clusters spread into comparatively large number of causal term clusters. This indicates that the documents inside the causal groups identified by our proposed framework are more causal at feature level than a grouping discovered by a similarity based baseline approach.

6.2 Lineage Formation

Figure 5 shows an example of a lineage identified by Algorithm 1 for a document in the IEEE Xplore library (Article Reference Number: 6729600). The branching factor we used for this lineage is $b = 4$ and the look-back threshold $r = 10$. The initial document for which the chain is formed embeds spatio-temporal data, social media, and recommender system in one paper. The causal paths shown in the lineage illustrates that the topic of this document was influenced by different areas of research including relational database with temporal aspects, association rule mining, network analysis, and viral marketing in social media. The chain discovers more causal articles than similar ones to form the lineage.

Overall, we evaluate the lineage in terms of diffusion coefficient (Equation 3) and citation overlap (Equation 4). Figure 6(left) compares average citation overlap scores at different levels of all the lineages detected for all documents using three approaches (1) our causality based approach, (2) similarity based approach, and (3) information genealogy based approach. The figure shows that the similarity based approach has the highest overlap with the citation lineage. This matches the fact that citations are generally outcomes

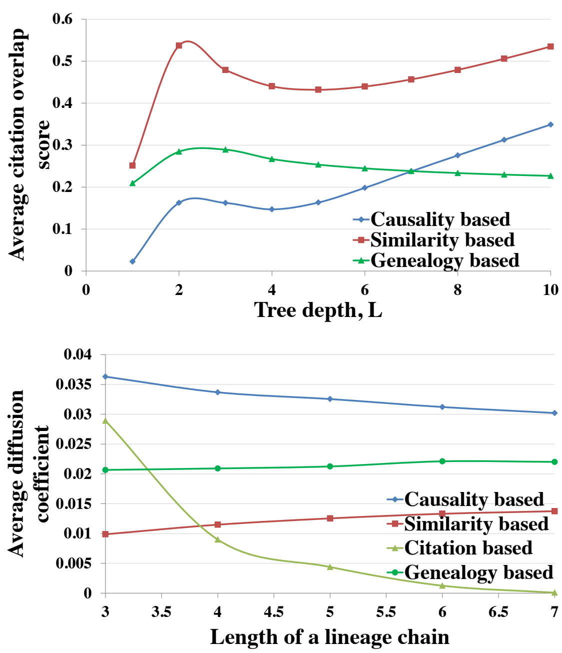

Figure 6: (left) Average citation overlap scores for three approaches: 1) our causality based approach, 2) similarity based lineage, and 3) information genealogy based approach. (right) Our causality based framework demonstrates the highest form of conceptual diffusion over time.

of keyword search resulting in similar topics. Our approach exhibits the least amount of overlaps with the citation network among three approaches. In contrast, Figure 6(right) shows that our approach has the highest average conceptual diffusion over time than all other methods, even when compared to the original citation network. This indicates that our approach has the ability to detect how the lineage drifted from another topic and formed the current literature, which other methods do not possess.

6.3 Analyzing Topical Causality

In this section we infer the notion of document-level causality to understand topic-level influential relationships in the literature. For the experiments in this section we used the documents published in ICDM conferences and workshops from 2001 to 2014, which is a subset of the IEEE dataset used in the preceding experiments. We apply LDA (Blei, Ng, and Jordan 2003) with 20 topics on this ICDM data set to get the distribution of topics in each document. We categorize a document to be under the topic which has the highest probability in that document.

If the documents under a particular topic are mostly caused by the other documents in that topic, the topic is supposed to depict more causal density (Equation 5). Figure 7(left) shows the causal density of different topics for $b = 5$. From the figure we can see that documents related to outliers detection (topic no. 8) and association rule and subgraph mining (topic no. 10) are mostly caused by the documents under the same topic. On the other hand, the research areas that inspired or were inspired by a variety of other fields should demonstrate strong causal relations across the topics. Figure 7(middle) shows the pairs of topics that have most causal influence calculated by Equation 6. Each edge in the figure is labeled with the value of $\phi(t_1, t_2)$. There are some areas in the literature that are more dominant than others and hence cause more documents within and across the topics. Figure 7(right) shows the top five dominant topics with their causal strengths computed by Equation 7.

Based on the causal relationships found among the topical groups of documents, our observation is that the topics with higher cross-topic causality has low causal density (e.g. topic number 5 and 18). Topic 10 is an exception in that sense, which is the reason behind its being the most dominant topic in the entire corpus.

7 Conclusion

In this paper we present a novel time series based representation for scientific articles that enables searching beyond mere content similarity. Though such representation is intuitive and has shown superior ability in identifying actual causal relationships between the documents, we are yet to comprehend all of the characteristics of these signals. One direction of our research is going towards having a better understanding of these document signals and applying them to identify even more subtle causalities. We also have shown a mechanism to find a causal chain of documents for any given document and then inferred those chains to topical groups of documents in order to see the causal relationships between

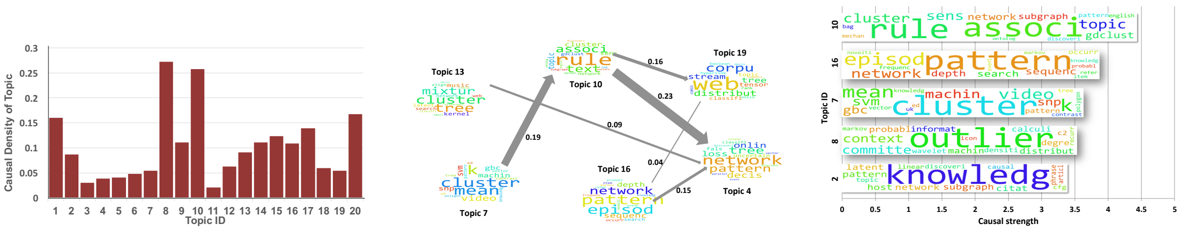

Figure 7: (left) Within-topic causality score of different topics. (middle) Topic pairs with most causal relation between them. (right) Topics with most causal strengths.

different topics. However, a direct approach towards understanding the topical evolution – how one concept begets other new ideas, gets merged with some other concepts, or simply vanishes away at some point – and the causal factors behind those behaviors of topics would be more interesting.

8 Acknowledgments

This work is supported in part by University Research Institute (URI, Office of Research and Sponsored Projects, UTEP) and U.S. Army Engineering Research and Development Center under Contract No. W9132V-15- C-0006.

References

AlSumait, L.; Barbará, D.; and Domeniconi, C. 2008. Online lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *ICDM'08*, 3–12.

Blei, D. M., and Lafferty, J. D. 2006. Dynamic topic models. In *ICML'06*, 113–120.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3:993–1022.

Caragea, C.; Bulgarov, F.; Godea, A.; and Gollapalli, S. D. 2014. Citation-enhanced keyphrase extraction from research papers: A supervised approach. In *EMNLP'14*.

Cheng, B. S., and Lai, T. W. 1997. An investigation of co-integration and causality between energy consumption and economic activity in taiwan. *Energy Economics* 19(4):435–444.

El-Arini, K., and Guestrin, C. 2011. Beyond keyword search: discovering relevant scientific literature. In *KDD'11*, 439–447.

Granger, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* 424–438.

Hasan, M. A.; Spangler, W. S.; Griffin, T.; and Alba, A. 2009. Coa: Finding novel patents through text analysis. In *KDD'09*, 1175–1184.

Hossain, M. S.; Butler, P.; Boedihardjo, A. P.; and Ramakrishnan, N. 2012. Storytelling in entity networks to support intelligence analysts. In *KDD'12*, 1375–1383.

Kataria, S.; Mitra, P.; and Bhatia, S. 2010. Utilizing context in generative bayesian models for linked corpus. In *AAAI'10*, volume 10, 1.

Mei, Q., and Zhai, C. 2005. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD'05*, 198–207.

Roebroeck, A.; Formisano, E.; and Goebel, R. 2005. Mapping directed influence over the brain using granger causality and fmri. *Neuroimage* 25(1):230–242.

Shaparenko, B., and Joachims, T. 2007. Information genealogy: uncovering the flow of ideas in non-hyperlinked document databases. In *KDD'07*, 619–628.

Teufel, S., and Moens, M. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics* 28(4):409–445.

Wang, X., and McCallum, A. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *KDD'06*, 424–433.