Weaving Information Propagation: Modeling The Way Information Spreads In Document Collections

Charles Huyghues-Despointes, Leila Khouas, Julien Velcin, and Sabine Loudcher

[1] Université de Lyon, Lyon 2, ERIC EA 3083, France,
charles.huyghues-despointes@univ-lyon2.fr
[2] Bertin IT, leila.khouas@bertin.fr
[3] Université de Lyon, Lyon 2, ERIC EA 3083, France,
julien.velcin@univ-lyon2.fr
[4] Université de Lyon, Lyon 2, ERIC EA 3083, France,
sabine.loudcher@univ-lyon2.fr

Abstract. Information usually spreads between people by the mean of textual documents. During such propagation, a piece of information can either remain the same or mutate. We propose to formulate information spread with a set of time-ordered document chains along which some information has likely been transmitted. This formulation is different from the usual graph view of a transmission process as it integrates a notion of lineage of the information. We also propose a way to construct a candidate set of document chains for the information propagation in a corpus of documents. We show that most of the chains have been judged as plausible by human experts.

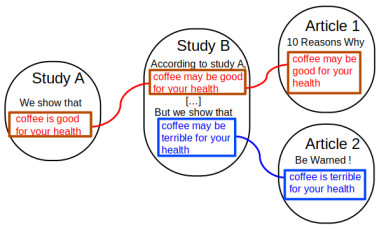
Keywords: information lineage, information propagation, document chains, textual data stream

1 Introduction

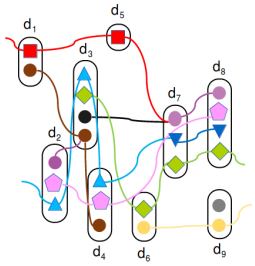
Internet came with a burst in document publication and accessibility. Nowadays, one can get far more news articles, videos or audio podcasts than one can digest on a daily basis. This has fostered new tool-assisted methods that take advantage of computer fast processing capabilities. Most of these documents convey information expressed in natural language. To analyze this information, the user mostly use search tools with a preconceived idea of what he is searching through a question or keywords. The user may need tools to understand how information propagates without using these preconceptions.

There are research efforts to study why and how information propagates in a corpus. Some works are interested on modeling the diffusion process between authors of documents in order to recover the propagation network as [1, 2]. Whereas these methods give us insights on the diffusion process between authors, they do

not express the diffusion between documents. The difficulty of uncovering the diffusion over documents is pointed in [3, 4] where authors try to trace a piece of information back to its primary source. Most of these models consider the diffusion as an information exchange [5] with no modification. However, during the information diffusion over documents, the specific context a piece of information appears in, may slightly make its meaning mutate, even when the information syntax does not evolve [6]. Moreover, those models expect pieces of information to be explicitly defined beforehand, which makes it difficult to express mutations. Furthermore, if two pieces of information share a similar context, as in Figure 1a, knowing their diffusion history could give us insights on how and why they differ. We propose to study how information propagates through textual document streams. We consider that a succession of documents, that we call a *chain*, is the basic structure for describing how information propagates. A set of chains is represented in Figure 1b.



(a) There is a lineage from Study A and Study B to Article 1, and a lineage from Study B to Article 2 only.



(b) The colored symbols stand for (up to now unknown) pieces of information.

We never know how the propagation phenomenon unfolds precisely. However, we consider that if information flows from a document to another, there should be some semantic similarity between these two documents. Given this assumption, we propose an algorithm to construct chains of similar documents as first candidates for propagation chains. Then, we present the results and our methodology to evaluate these chains. We show that human experts reach a consensus on what is a plausible chain of propagation and what is not, and that our calculated chains match with this evaluation. We conclude in giving several applications that can be built on top of this general chain propagation model.

2 Modeling Information Propagation With Chains

We do not formally define what one piece of information is. It can be expressible facts, ideas, opinions, reasoning or even sensations, but it also may be some complex discrete representation of information. We denote K the set of all pieces of information, and D a corpus of documents. An interesting property of K

elements is that they may be more or less similar. Statements like "It may rain today" and "It will rain today" are not exactly the same, still they share some semantic. We set $sem_K : K \times K \to [0, 1]$ a semantic similarity metrics over K.

During propagation between the two documents d_i and d_j, a piece of information k_0 of d_i may mutate, resulting in a possibly different k_1 in d_j. We note such mutation as a pair (k_0, k_1). We can express the propagation event between d_i and d_j by enumerating all the mutations that occur between them, noted as $M_{i,j}$. We define a propagation event e as a triplet $e = (d_i, d_j, M_{i,j})$ for $d_i, d_j \in D$. There exists a non-trivial threshold ϵ such that: if $(d_i, d_j, M_{i,j})$ is a propagation event then $(k_i, k_j) \in M_{i,j} \implies sem_k(k_i, k_j) \geq 1 - \epsilon$.

This propagation model can express historical modifications of information. Given two propagation events $(d_1, d_2, \{(k_1, k_{1'})\})$ and $(d_2, d_4, \{(k_{1'}, k_{1''})\})$, we can flow back the origin of $k_{1''}$ in d_2, as k_1 in d_1. Thus, information has propagated along the path $d_1 d_2 d_4$. Such path of documents is what we call a *propagation chain*. Given P a set of propagation events over documents, a chronologically ordered sequence of documents $c = d_0 d_1 \dots d_n$ is a propagation chain if:

$$\forall i \in \{1, 2, \dots, n\}, \exists\ e_i, e_{i+1} \in P / \begin{cases} e_i = (d_{i-1}, d_i, M_{i-1,i}), \\ e_{i+1} = (d_i, d_{i+1}, M_{i,i+1}) \\ \exists (k, k') \in M_{i-1,i} \wedge \exists (k', k'') \in M_{i,i+1} \end{cases}$$

We denote the set of propagation chains by T, which stands for trajectory. Conceptually, each transition of a propagation chain must keep a common endpoint (document d_i) and a common semantic endpoint (piece of information k'). Note that T has the following property: if $c = d_0 d_1 \dots d_n$ is a propagation chain of T, then $\forall 1 \leq i < j \leq n, c[i, j] = d_i \dots d_j$ is also a propagation chain of T. Furthermore, we say that $c[i, j]$ is a sub-chain of c.

3 An Approach To Propagation Chain Approximation

In this section, we assume known a corpus D. We denoted by T_D the chains only composed of documents from D. We do not explicitly know the information pieces contained in documents. Instead, we have a similarity function sim between documents. In order to compute a good approximation of T_D, we construct coherence metrics for propagation chains based on that similarity. Then, we compute the chains that satisfy this metric up to a given threshold.

It seems reasonable that most of the propagation chains should be coherent chains. We model the coherence using a metric, denoted by coh that assigns a number between 0 and 1 to a chain. We say that a chain c is *coherent* if $coh(c) > 1 - \epsilon$, with ϵ a given coherence threshold. In order to construct every document chain satisfying our coherence criterion, we make use of the property stipulating that if c is a propagation chain, then every sub-chain c is also a propagation chain. Extending this proposition to coherent chains implies that every sub-chain of c must satisfy our coherence criterion. This allows us to use a dynamic programming approach. We define $FinishIn(d)$ as the set of coherent

chains finishing by d. In order to finish in d, a document chain must be the concatenation (operator .) of a chain $d'd$ and at most one chain from $FinishIn(d')$. We can then construct the Candidates for d and the coherent chains T_{coh} :

$$Candidates(d) = \bigcup_{d'/coh(d'd)>1-\epsilon} \{c.d/c \in FinishIn(d') \cup \{d'\}\}$$

$$FinishIn(d) = \{c \in Candidates(d)/coh(c) > 1 - \epsilon\}$$

$$T_{coh} = \bigcup_{d \in D} FinishIn(d)$$

We can solve this problem using a bottom-up strategy, starting from oldest to newer documents, since a chain always respects the publication chronology. This approach has some complexity issues due to the potentially exponential number of coherent chains. For that purpose, we introduce a constant safeguard threshold on the maximal number of coherent chains finishing in every document. This ensures a linear memory complexity.

4 Experiments

We sampled two datasets of 150 documents. The first one comes from the Citation Network Dataset V1 AMINER[5] constructed by [7]. This is mostly abstracts from scientific papers extracted from ACM and DBLP. Our second dataset is sampled from the full set of news articles posted on the US version of the Huffington Post from 1st July to 30th November 2016. For document similarity, we used a TFIDF cosinus similarity where we enriched the bag of words with n-grams of size 2 to 4. For chain coherence, we used the minimal similarity and the arithmetic mean of similarity between document pairs inside the chain. Three different coherence thresholds are considered: 10%, 20% and 50%. It yields six coherent chain sets for each dataset (81 and 149 chains to evaluate, respectively).

For the sake of annotation, we gathered experts who have a professional level in English and a good knowledge of reading scientific papers. We spread chains between experts such that every chain is at least evaluated twice and two experts are not exposed to the same succession of chains. For each document chain, an expert evaluates links between two successive documents (determining if there is a strong, weak or no semantic link). Then, he evaluates sub-chains from the first document to every other determining if it is strongly, weakly or not plausible that some information propagated.

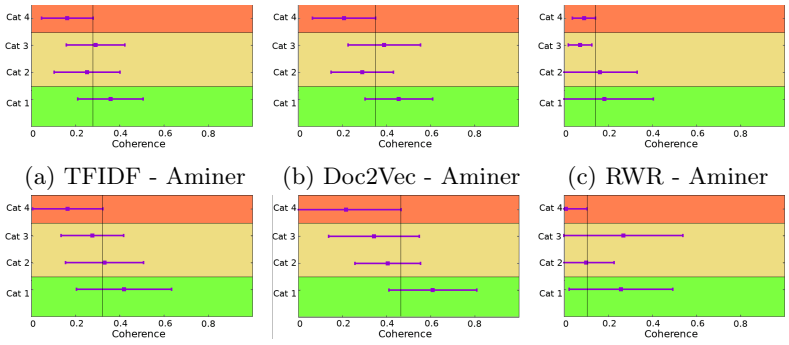
Agreement ratio is computed as the proportion of paired evaluations with the same conclusion over the total number of paired evaluations. We speak of presence if the chain or the link is annotated as strongly or weakly plausible. Experts agreed on more than 70% cases on the semantic link presence for both datasets and it exceeds 80% for the presence of a plausible chain. This result reinforces the intuition that it is easier to reach a consensus when we have more

[5] The full dataset can be obtained at: https://aminer.org/citation

context. This shows that experts can have an intuition of chain plausibility with consistency, which means that detecting coherence in a chain is a feasible task.

Now, we label the annotation for each chain with five categories for both the semantic link task and the chain plausibility task. Each category represents the majority of annotations. **Category 1** stands for **strong intensity**. **Category 2** is for **weak intensity**. **Category 3** stands for **presence** but with no intensity agreement. **Category 4** is for **absence**. **Category 5** is finally for the case where there is no majority agreement. Results on AMINER are good with nearly 70% of plausible chains (Cat. 1, 2 and 3) and 75% of linked chains. On the other hand, Huffpost results are much weaker. 64% of evaluated links are judged non-existent and 75% of chains are judged non plausible. We will show that these non relevant chains come from a low coherence threshold.

Annotated chains serve as a ground truth allowing us to compare different coherence criterion. Studied similarity metrics are: TFIDF with a cosine similarity; Doc2Vec with a cosine similarity, with vectors of size 20 learnt on the dataset; RWR: a random-walk-based approach [8], parametrized with a 99% restart probability. For the coherence, we use the arithmetic mean coherence which is computed as the mean similarity between document pairs from the same chain. We consider the coherence as a good discrimination function if there is a small or even null intersection over coherence range (defined as the mean plus or minus the standard deviation) for strongly plausible and non-plausible chain categories. These intervals are presented in Fig. 2. On both datasets, the Doc2Vec-based coherence is the most discriminating metric. Generally, strongly and weakly plausible chains have higher coherence than non-plausible ones. For Huffpost, the results explain the proportion of bad chains observed. A typical HuffPost bad chain has a TFIDF based coherence under 0.2. It means that those chains mostly come from the trajectory computed with a 0.1 coherence threshold. These coherence metrics prove that capturing human judgment over document chains is partly possible by using well-known similarities.



(a) TFIDF - Aminer (b) Doc2Vec - Aminer (c) RWR - Aminer

(d) TFIDF - HuffPost (e) Doc2Vec - HuffPost (f) RWR - HuffPost

Fig. 2: Mean and standard deviation for different document similarities by annotation category. A vertical line marks the absence category (Cat 4) overlap.

5 Conclusion

Considering the information propagation through trajectories over a textual document network is a novel idea. An important advantage over other methods based on the propagation graph lies in a better understanding of the history of information propagation. We proposed an approximation of trajectory by coherent chains that can be solved through dynamic programming. To qualify the computed chains, we realised a human evaluation campaign. This campaign had two benefits. First, we saw that human evaluations were consistent between themselves, which suggests that recognizing a plausible propagation chain is feasible. Second, we used evaluations as a ground truth for testing different coherence criterion. We saw that criterion based on well-known metrics succeed in capturing human judgment. We consider this result as a first proof that this task may be solved using an automatic process.

For future work, we plan to overcome the necessity of guessing a correct coherence threshold. We also plan to automatically identify the pieces of information that propagate along the chains. We foresight multiple use cases for a good trajectory approximation. One use case is to easily navigate in the document space for an analyst user by following an interesting subject flow, or to give him a good understanding of the information flow by summing up chains into a metromap of information. It may also be an interesting framework to study how information pieces interact with each other along the chains.

References

1. Gomez-Rodriguez, M., Balduzzi, D., Schölkopf, B.: Uncovering the temporal dynamics of diffusion networks. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Washington, USA. (2011) 561–568
2. Zarezade, A., Khodadadi, A., Farajtabar, M., Rabiee, H.R., Zha, H.: Correlated cascades: Compete or cooperate. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA. (2017) 238–244
3. Pinto, P.C., Thiran, P., Vetterli, M.: Locating the source of diffusion in large-scale networks. Phys. Rev. Lett. **109** (Aug 2012) 068702
4. Farajtabar, M., Gomez-Rodriguez, M., Zamani, M., Du, N., Zha, H., Song, L.: Back to the past: Source identification in diffusion networks from partially observed cascades. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA. (2015)
5. Zafarani, R., Ali Abbasi, M., Liu, H.: Social Media Mining, An Introduction. Cambridge University Press (2014)
6. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the Dynamics of the News Cycle. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '09, USA, ACM (2009) 497–506
7. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: Extraction and mining of academic social networks. In: KDD'08. (2008) 990–998
8. Shahaf, D., Guestrin, C.: Connecting the dots between news articles. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '10, New York, USA, ACM (2010) 623–632