Extraction de chaînes cohérentes en vue de reconstuire la Trajectoire de l'information

Charles Huyghues-Despointes[*,**], Leila Khouas[**], Julien Velcin[*] et Sabine Loudcher[*]

[*]Laboratoire ERIC
[**]Bertin IT

Résumé. Sur Internet, l'information se propage en particulier au travers des documents textuels. Cette propagation soulève de nombreux défis : identifier une information, suivre son évolution dans le temps, comprendre les mécanismes qui régissent sa propagation, etc. Étant donné un document parmi un grand corpus dans lequel de nombreuses informations circulent, pouvons-nous retrouver les chemins empruntés par l'information pour arriver à ce document ? Nous proposons de définir la notion de trajectoire comme l'ensemble des chemins le long desquels de l'information s'est propagée et nous proposons une méthode pour l'estimer. Nous avons mis en œuvre une évaluation humaine pour juger de la qualité des chemins calculés. Nous montrons que les évaluations concordent la plupart du temps et que notre algorithme est efficace pour retrouver les bons chemins.

1 Introduction

L'information se propage. Lorsqu'elle est reçue, une information est ingérée, nuancée, et reformulée pour être à nouveau transmise. Cette propagation se déroule à tous les niveaux de communication : lors d'une conversation, à la radio, à la télévision, mais aussi lorsque nous publions du contenu, par exemple sur Internet. Les documents que nous partageons, contiennent de multiples informations provenant d'autres documents, et qui seront, en partie, reprises dans le futur. Ainsi les informations présentes dans un document ont une histoire. Ce sont des suites d'événements de propagations qui les ont conduites à être présentes dans ce document. Nous appelons l'ensemble de ces lignées, pour chaque document, la Trajectoire de l'information.

Lorsqu'une information se propage elle est sujette à des modifications, dans sa forme ou dans son fond. Certains travaux se sont intéressés à la traque de ces changements, comme par exemple Leskovec et al. (2009). Cependant, après de nombreuses mutations, il peut être difficile de trouver le lien entre l'information de départ et l'information actuelle, comme le soulignent des travaux cherchant à retrouver les sources d'une information, comme par exemple Farajtabar et al. (2015).

Nous proposons d'estimer la Trajectoire de l'information en calculant des chaînes de documents textuels le long desquelles il est plausible que de l'information se soit propagée. Pour ce faire nous n'explicitons pas l'information qui circule le long de la chaîne, mais nous intéressons à la manière dont se comportent les documents entre eux au sein de la chaîne. Aborder

```
1 Donald Trump Is Also an Outlier in Political Science
2 Donald Trump Is Forcing Ted Cruz to Rewrite His Playbook
3 The Republican Establishment Is Losing at Its Own Game
4 Ted Cruz and Allies Work to Halt Donald Trump's Gains
```

FIG. 1: Chaîne de propagation plausible tirée d'articles du New York Times (titres affichés)

FIG. 2: Deux trajectoires de mêmes support. {[ACD], [BCE]} et {[ACD],[ACE],[BD]}.

le problème de la propagation d'information à partir d'une telle structuration du corpus n'a pas, à notre connaissance, été traité dans la littérature. Nous inscrivons notre travail comme un premier pas dans cette direction. Nous commençons par détailler notre objectif et notre approche en première section. Nous avons mené une campagne d'évaluation auprès d'experts qui suggère que la plausibilité d'une chaîne est estimable et que notre approche, bien que simple, donne de bons résultats. Notre protocole d'évaluation et nos résultats sont discutés en deuxième partie. Nous concluons sur les perspectives d'utilisation de cet ensemble de chaînes et nos idées pour améliorer et approfondir notre approche.

2 Extraction des chaînes de propagation

Le contexte du problème est le suivant : nous analysons un ensemble de documents textuels (un corpus) dont nous connaissons certaines méta-données, comme la date de publication, les auteurs, etc. Notre première hypothèse est l'existence d'un phénomène de propagation de l'information : durant le processus de création des documents, les auteurs récupèrent, interprètent et reformulent différentes informations issues de documents antérieurs du corpus (ou d'ailleurs). Notre seconde hypothèse est qu'une information qui a muté garde un lien sémantique fort avec l'information dont elle dérive.

On appelle **chaîne de propagation** une chaîne de documents le long de laquelle au moins une information s'est propagée au sens évoqué ci-dessus. Nous appelons **trajectoire** un ensemble de chaînes de documents. On dit qu'une chaîne de documents est une **chaîne de propagation plausible** si des évaluateurs humains s'accordent pour dire qu'il a pu y avoir une propagation d'information le long de cette chaîne. Un exemple de chaîne de propagation plausible est donné en Fig. 1. Une trajectoire n'est un graphe sur les documents. La Fig 2 montre deux trajectoires différentes utilisant les mêmes arêtes. Notre objectif est le suivant : calculer une trajectoire contenant le plus de chaînes de propagation plausibles, c'est-à-dire cohérentes, et le moins de chaînes non plausibles.

Notre approche consiste à parcourir toutes les chaînes possibles et à sélectionner celles qui satisfont un certain critère de cohérence. Toutes les chaînes ne sont pas possibles, en particulier elles doivent satisfaire deux propriétés. La première est une propriété de croissance : une chaîne $ABCD$ est une chaîne de propagation à condition que CD le soit aussi. Sinon, cela veut dire qu'une information circule le long de $ABCD$ sans qu'aucune ne circule le long de CD. Ainsi,

si CD ne satisfait pas notre critère de cohérence, nous n'explorons pas les chaînes qui passent par CD. La seconde exploite la date de publication des documents. Une information se propage toujours du document le plus ancien vers le document le plus récent.

Nous procédons de la manière suivante : nous calculons pour chaque document D les chaînes qui finissent en D, que nous notons $FinishIn(D)$. Pour cela, nous calculons l'ensemble des chaînes candidates pour D, que nous notons $Candidates(D)$. Les chaînes candidates pour D sont toutes les chaînes formées de documents publiés avant D. Étant donnée notre propriété de croissance, nous parcourons les chaînes qui finissent en C à la condition que la chaîne CD satisfasse notre critère de cohérence. Une fois tous les candidats accumulés, les chaînes qui finissent en D sont le résultat de notre stratégie de sélection $select$. La trajectoire calculée T est l'union de toutes les chaînes calculées. Le pseudo-code de l'algorithme est donné en Algorithme 1.

Data : un corpus de document $Corpus$, une stratégie de sélection $select$
Result : T l'ensemble des chaînes calculées
$Treated \leftarrow \emptyset$;
$T \leftarrow \emptyset$;
for $D \in Corpus$ *par date de publication croissante* **do**
  $FinishIn(D) \leftarrow \emptyset$;
  $Candidates(D) \leftarrow \{D\}$;
  for $C \in Treated$ *vérifiant* $date(C) < date(D)$ *et* $select(\{CD\}) \neq \emptyset$ **do**
    $Candidates(D) \leftarrow Candidates(D) \cup \{chain.D, chain \in FinishIn(C)\}$;
  end
  $FinishIn(D) \leftarrow select(Candidates(d))$;
  $T \leftarrow T \cup FinishIn(D)$;
  $Treated \leftarrow Treated \cup \{D\}$
end
return T

Algorithme 1 : Calcul d'une trajectoire de l'information

Notre stratégie de sélection est la suivante : nous définissons la mesure d'attachement d'un document à une chaîne comme une mesure de la vraisemblance de l'ajout du document à la fin de la chaîne. Dans nos expériences, nous avons donné à l'attachement du document D à la chaîne ABC la forme suivante : $attach(D, ABC) = F(sim(A, D), sim(B, D), sim(C, D))$ où sim est une fonction de similarité sémantique entre documents. F peut être une fonction simple comme le minimum ou une moyenne. Nous sélectionnons les k chaînes maximales selon l'attachement pour le document actuel avec comme contrainte que l'attachement doit être supérieur à un seuil de cohérence.

3 Expérimentations

L'estimation de la Trajectoire comme nous le proposons étant un problème neuf à notre connaissance, nous nous sommes tournés vers l'évaluation humaine de manière à construire des jeux de données annotés. Nous avons pris deux jeux de données anglophones. Le premier est le Citation Network Dataset V1 d'AMINER [1] construit par Tang et al. (2008). Il est composé de résumés de papiers scientifiques extraits de collections comme ACM et DBLP. Notre second jeu de données correspond à l'ensemble des articles du Huffington Post US sur la période du 1er juillet au 30 novembre 2016. Les jeux contiennent respectivement 629 814 et 49 648 documents.

Nous avons créé deux jeux de données dérivés contenant moins de documents pour avoir un nombre de chaînes à évaluer raisonnable. Nous avons choisi de sélectionner 150 résumés

1. Le jeu AMINER est disponible ici : `https://aminer.org/citation`

au hasard pour les deux corpus. Cependant pour le Huffington Post, nous avons été plus précis. Nous avons enlevé les articles contenant le mot "Trump" très représenté dans le jeu (> 11 000 documents). Nous n'avons gardé que les articles entre 100 et 3000 signes pour ne pas perdre l'attention de l'évaluateur dans des articles trop longs à lire ou qui présentent trop peu de contexte. Nous avons créé nos trajectoires à partir d'une similarité cosinus sur les vecteurs TFIDF des documents. Nous avons construit six trajectoires en faisant varier la mesure d'attachement F d'une part (la moyenne arithmétique ou le minimum) et le seuil d'admissibilité d'autre part (parmi les valeurs 0,1 ou 0,2 ou 0,5). Nous avons réuni ces trajectoires pour chaque jeu de données. Cela nous donne deux ensembles de chaînes à évaluer.

Dans le cas d'une évaluation humaine, l'expertise des évaluateurs entre en jeu. Nous avons demandé à quatre chercheurs en informatique d'annoter les chaînes que nous avons calculées. Ils sont habitués à lire des documents tels que ceux d'AMINER. Les articles du Huffington Post sont destinés à un lectorat étendu et nous n'avons pas remis en cause la capacité de nos participants à les comprendre et à les mettre en contexte. La démarche de l'évaluateur est la suivante : d'abord, l'évaluateur doit prendre connaissance du contexte de la chaîne. Ensuite, il lit le premier document de la chaîne (le plus ancien). Puis, chacun des documents suivant lui est proposé en succession. À partir de là, il doit pour chaque déterminer s'il y a un lien sémantique fort ou faible avec le document précédent et s'il est fortement/faiblement/non plausible que de l'information se soit propagée du premier document jusqu'à celui-ci.

Nous donnons dans la Tab. 1a le ratio d'accord des participants pour l'évaluation des liens directs et celle de l'attachement pour les chaînes d'au moins trois documents. Nous séparons les résultats en deux, selon qu'on considère l'intensité du lien ou juste son existence. Pour les liens directs, les évaluateurs sont d'accord dans au moins 70 % des cas sur les deux jeux de données et dans au moins 80 % des cas (sauf pour l'intensité sur AMINER) pour l'attachement. Cela renforce l'intuition que l'évaluation est plus facile quand le contexte est plus riche. Ces deux résultats montrent que les humains arrivent à évaluer la cohérence des chaînes de documents avec consistance. Ceci nous conforte dans l'idée que le problème que nous traitons est bien posé.

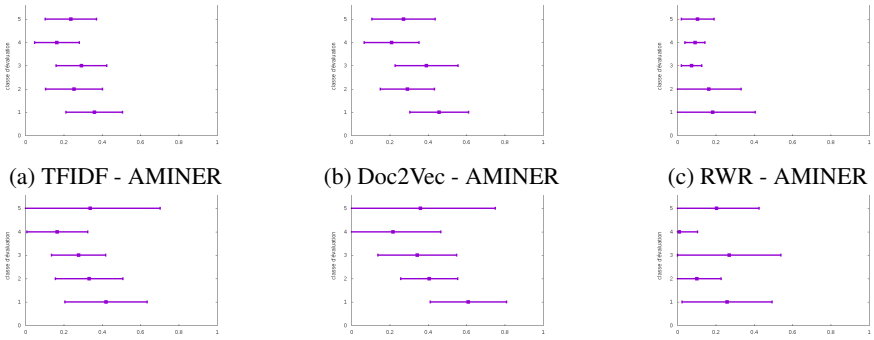
(a) Accord inter-évaluateurs

Objet évalué	propriété	AMINER	HuffPost
Lien avec le	évaluations	81	149
doc précédent	Fort/Faible/Non	68.09%	77.27%
	Lien/Non	76.60%	79.55%
Attachement avec	évaluations	66	107
la chaîne	Fort/Faible/Non	57.89%	83.70%
(chaîne de taille > 2)	Lien/Non	80.70%	85.87%

(b) répartitions des évaluations (en %)

	Catégorie	1	2	3	4	5
Lien direct	AMINER	40.7	23.5	4.9	17.3	13.6
	HuffPost	18.8	10.7	1.3	63.8	5.4
Attachement	AMINER	34.8	19.7	19.7	9.1	16.7
	HuffPost	7.5	4.7	1.9	74.7	11.2

Nous choisissons de répartir nos évaluations en cinq catégories selon l'accord des évaluateurs. La majorité a jugé qu'il y avait : un lien fort (Catégorie 1), un lien faible (Catégorie 2), un lien sans trancher sur son intensité (Catégorie 3), une absence de lien (Catégorie 4). Il y a une catégorie 5 qui est le cas ou la majorité n'est pas atteinte. La répartition est donnée en Tab. 1b. Nous remarquons que les résultats sont très bons pour AMINER avec seulement 9 % de non-attachement. A contrario, les chaînes sur le HuffPost sont globalement mauvaises à la fois pour le lien direct (64 %) et pour les attachements (75 %). Pour comprendre ce résultat, nous devons nous rappeler comment a été créé l'ensemble de chaînes que nous évaluons. Il s'agit de l'union de plusieurs trajectoires, parmi lesquelles deux trajectoires calculées avec un seuil d'admissibilité de 0,1. Nous montrons plus loin que les mauvaises chaînes proviennent



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

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

FIG. 3: Moyenne et écart-type des mesures d'attachement sur les chaînes par catégorie

de ces trajectoires. Cette différence entre AMINER et HuffPost pose le problème du choix du seuil d'admissibilité : le seuil idéal dépend du jeu de données.

Maintenant nous avons des chaînes annotées, nous disposons d'une vérité terrain qui nous permet de rechercher un critère de cohérence plus pertinent. Nous proposons d'étudier qualitativement des fonctions d'attachement basées sur d'autres similarités. En plus du TFIDF, nous étudions : Une similarité basée sur Doc2Vec (Le et Mikolov, 2014) (avec un espace de taille 20). Nous considérons aussi une similarité calculée par marche aléatoire avec retour, développée par Shahaf et Guestrin (2010), que nous nommons RWR, paramétrée avec une probabilité de retour de 99 %. Toutes les similarités sont entraînées sur l'intégralité des documents des corpus créés. Nous définissons une mesure d'attachement par moyenne arithmétique pour chacune de ces mesures. Pour chaque catégorie de chaînes annotées, nous calculons la moyenne et l'écart-type de l'attachement présenté sous forme d'intervalles en Fig. 3. Nous remarquons que les trois mesures attribuent un meilleur score aux chaînes jugées liées qu'aux chaînes jugées non liées. En particulier Doc2Vec semble être la mesure qui dissocie le mieux les chaînes fortement liées des chaînes non liées. Ceci montre qu'il est possible de capturer au moins en partie le jugement humain sur les chaînes avec des mesures bien connues. Si l'évaluation humaine montrait que la tâche est réalisable par des experts, celle-ci renforce notre intuition que la tâche est également réalisable par une machine.

4 Conclusion

Calculer des approximations de la Trajectoire est un problème encore ouvert. Nous avons proposé un cadre pour le formuler ainsi qu'une approche gloutonne qui calcule des chaînes de proche en proche. Dans le but de qualifier ces chaînes, nous avons mené une campagne d'évaluation humaine. Le bénéfice a été double : d'une part, nous avons vu que les évaluations humaines étaient consistantes entre elles, ce qui nous conforte dans l'idée que le problème est bien posé puisque la tâche est réalisable par l'humain. D'autre part, nous nous sommes servi de ces évaluations comme d'une vérité terrain pour tester différents critères de cohérence. Nous avons vu que ces critères réussissent à capturer les jugements humains. Nous interprétons ce résultat comme une première preuve que la tâche est aussi réalisable de manière automatique.

Plusieurs axes d'améliorations sont envisagés. En particulier, Nous comptons former un critère de cohérence plus performant encore. Pour cela, nous prévoyons une nouvelle campagne d'évaluation avec un nombre plus élevé de participants, ce qui aura aussi pour effet de consolider ou nuancer nos premiers résultats. Nous souhaitons également chercher de nouvelles façons de créer nos chaînes, par exemple en utilisant des méthodes probabilistes qui tireraient un ensemble de chaînes dont la cohérence serait élevée.

Une fois des trajectoires fiables calculées automatiquement, nous pouvons explorer leur utilisation dans plusieurs cas d'exploitation. Le but de la trajectoire est d'isoler les chaînes de propagation, aussi une volonté naturelle serait d'extraire les informations qui se propagent le long de chaque chaîne, mais aussi étudier la manière dont ses informations interagissent entre elles le long des chaînes. On peut aussi plonger les chaînes dans l'espace des auteurs afin d'étudier la manière dont ces derniers relaient l'information. Enfin, nous nous posons la question de la synthèse et de la visualisation des chaînes elles-mêmes. Cela peut être la constitution d'un résumé de la propagation de l'information, une piste prometteuse en ce sens réside dans les travaux menés par Shahaf et al. (2013).

Références

Farajtabar, M., M. Gomez-Rodriguez, M. Zamani, N. Du, H. Zha, et L. Song (2015). 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*.

Le, Q. V. et T. Mikolov (2014). Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*, pp. 1188–1196.

Leskovec, J., L. Backstrom, et J. Kleinberg (2009). 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, pp. 497–506. ACM.

Shahaf, D. et C. Guestrin (2010). 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, NY, USA, pp. 623–632. ACM.

Shahaf, D., C. Guestrin, et E. Horvitz (2013). "metro maps of information" by dafna shahaf, carlos guestrin and eric horvitz, with ching-man au yeung as coordinator. *SIGWEB Newsletter 2013*(Spring), 4 :1–4 :9.

Tang, J., J. Zhang, L. Yao, J. Li, L. Zhang, et Z. Su (2008). Arnetminer : Extraction and mining of academic social networks. In *KDD'08*, pp. 990–998.

Summary

We propose the notion of Trajectory as the set of paths along which some information spread and we show an algorithm for approximate it. We show that humans evaluations mostly match and that our algorithm finds good paths effectively.