Spatial Textual Representation (STR) ou comment représenter la spatialité des donnnées textuelles

Jacques Fize, Mathieu Roche, Maguelonne Teisseire

▶ **To cite this version:**

Jacques Fize, Mathieu Roche, Maguelonne Teisseire. Spatial Textual Representation (STR) ou comment représenter la spatialité des donnnées textuelles. Spatial Analysis and GEOmatics 2017, INSA de rouen, Nov 2017, Rouen, France. pp.14. hal-01643368

HAL Id: hal-01643368
https://hal.science/hal-01643368

Submitted on 21 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Spatial Textual Representation (STR) ou comment représenter la spatialité des données textuelles

Jacques Fize[*], **Mathieu Roche**[*], **Maguelonne Teisseire**[*]

** UMR 9000 TETIS, Cirad, Irstea, CNRS, AgroparisTech, Univ. Montpellier Maison de la Télédétection, Montpellier, France*
{firstname}.{lastname}@teledetection.fr

RÉSUMÉ. *De nos jours, la majorité des systèmes de recherche d'information (RI) basent leurs algorithmes de recherche et d'indexation de document sur leur dimension thématique. Dans le cadre de l'étude de la mise en correspondance de données textuelles hétérogènes, nous proposons une approche permettant d'exploiter la dimension spatiale des données textuelles à travers une structure appelée STR ou Spatial Textual Representation. STR est une structure de graphe qui permet de représenter la configuration spatiale d'un document à l'aide d'informations du texte et de sources externes (gazetier). Pour créer ces graphes, nous extrayons tout d'abord les entités spatiales. Puis, nous relions les entités spatiales selon différentes relations (adjacence, inclusion, ...). Enfin, nous évaluons plusieurs mesures de similarités, propres aux graphes, sur la structure proposée. Les expérimentations sont menées sur 3 corpus distincts.*

ABSTRACT. *Nowadays, many Information Retrieval (IR) systems base their research and indexation algorithm on the thematic dimension of the text. As part of the study of matching heterogeneous textual documents, we investigate the spatial dimension of textual data using an original approach called STR. STR or Spatial Textual Representation is a graph structure which allows to represent spatial configuration in text using intern and external sources of information. In order to create such graph, we extract spatial entities in documents. Then we link the spatial entities using different relationships (i.e neighbor, inclusion, ...). Finally, we evaluate different similarity measures related to graph on the proposed structure. The experiments were realized using 3 corpora.*

MOTS-CLÉS : *Recherche d'information, Fouille de données, Données spatiales*

KEYWORDS: *Information Retrieval, Data-mining, Spatial data*

1. Introduction

De nos jours, le volume de données présentes dans un contexte BigData ouvre de nouvelles opportunités et de nouveaux défis scientifiques pour gérer, stocker et exploiter ces données. Dans ce cadre, la mise en correspondance de données textuelles hétérogènes restent un défi pour la communauté scientifique. Il s'agit d'associer des documents qui possèdent des similarités et/ou complémentarités selon différentes dimensions, comme la thématique, la spatialité, la temporalité, etc. Pour cela, deux composantes sont nécessaires : (1) une représentation commune des données selon ces dimensions et (2) différentes mesures (Lin *et al.*, 1998) pour quantifier la similarité.

Actuellement, la majorité des méthodes de mise en correspondance base leurs algorithmes de recherche et d'indexation sur la dimension thématique des données textuelles. On peut mentionner l'approche sac de mots, ou *bag-of-words* (Salton, Mcgill, 1986), des modèles plus complexes comme LDA (Blei *et al.*, 2003) ou plus récemment, du *document embedding* (Le, Mikolov, 2014 ; Ye *et al.*, 2016). Même si ces techniques et modèles sont efficaces, ils se focalisent principalement sur la fréquence/co-occurrence des mots.

Dans cet article, nous nous focalisons sur l'exploitation de la dimension spatiale des textes. Depuis des années, l'analyse de la spatialité dans les documents textuelles intéresse les chercheurs de différents domaines. Par exemple, dans des domaines tels que l'épidémiologie, où l'information utilisée provient à 60% de données textuelles. Dans ce contexte, la recherche de similarité spatiale entre documents (presse locale, revue scientifique, etc.) est importante voir cruciale. On définit la spatialité dans un document selon deux types d'informations : les entités spatiales et les relations qu'elles entretiennent. Nous proposons de les prendre en compte par une représentation de type graphe, nommée Spatial Textual Representation (STR), qui s'inspire des QCN (Qualitative Constraint Network)(Ligozat, 2013).

La suite de cet article est organisée de la façon suivante. Dans la deuxième section, nous présentons un état de l'art des travaux connexes. Puis, dans la section 3, nous présentons la STR, son processus de création ainsi que différentes mesures de similarités. Enfin, les expérimentations sont détaillées dans la section 4, avant de conclure en section 5.

2. Travaux Connexes

La représentation de l'information spatiale a retenu l'attention de la communauté scientifique ces dernières années. Nous pouvons citer les travaux tels que (Belouaer *et al.*, 2016) sur l'analyse de trajectoire dans les textes, (Prudhomme *et al.*, 2017) (Gaio *et al.*, 2012) sur la création d'ontologies spatiales à partir de diverses sources de données (bases de données, textes, ...). Suivant une pers-

pective de mise en correspondance de données textuelles, nous avons choisi de présenter les travaux suivants.

La recherche de similarité entre des documents implique la création d'une représentation commune. Dans le cas de la spatialité, on parle de Geographical Information Retrieval (GIR), dédiée à l'indexation et l'élaboration de systèmes de recherches de médias dans leur dimension spatiale. Parmi les différentes méthodes proposées, il existe le *geocoding* qui désigne le processus qui assigne une ou plusieurs coordonnées (longitude, latitude) à un document. Différents travaux tels que ceux de (Woodru, Plaunt, 1994 ; Amitay *et al.*, 2004) proposent une méthode d'extraction d'un contexte géographique en utilisant la géométrie des différentes entités spatiales. D'autres travaux (Wing, Baldridge, 2011) s'inspirent des techniques de "language modeling", pour calculer la probabilité d'appartenance d'un document à un zone géographique.

Dans nos travaux, nous avons choisi de nous intéresser aux Qualitative Constraint Network (Ligozat, 2013). Par exemple, dans (Wallgrün *et al.*, 2010), les auteurs proposent une méthode de recherche d'information spatiale basée sur une description utilisateur ou *sketch*. Pour cela, la requête utilisateur et la base de données sont transformées en QCN. Dans (Al-Salman *et al.*, 2012), le principe est similaire à (Wallgrün *et al.*, 2010), excepté que la base de données est relationnelle (POSTGIS). Cette situation implique que pour rechercher la ou les informations correspondantes au QCN extrait du *sketch*, il faut transformer celui-ci en requête SQL.

En faisant le bilan des approches présentées, nous pouvons souligner que les méthodes de geocoding ne prennent en compte que les caractéristiques de geolocalisation des différents objets du document. Les QCN proposent quant à eux une représentation de la spatialité différentes prenant en compte les relations entre les entités spatiales. Les applications et approches proposées dans la littérature autour des QCN, se trouvent généralement dans l'interprétation du langage naturel sous forme de requête ou de description d'un lieu. Dans notre approche, nous proposons d'utiliser et d'étendre l'information contenue dans ces représentations pour l'indexation de document dans des systèmes de Recherche d'Information (RI).

3. STR : Spatial Textual Representation

Une STR, ou Spatial Textual Representation, est une représentation de la spatialité dans des données textuelles, utilisant une structure de graphe basée sur les QCN, ou Qualitative Constraint Network. Elle est générée à partir de données externes aux données textuelles (gazetier) puis enrichie à l'aide de données internes.

Nous proposons la construction des STR en 3 étapes (voir Figure 1). Premièrement, les entités spatiales sont identifiées à l'aide d'un module d'extraction

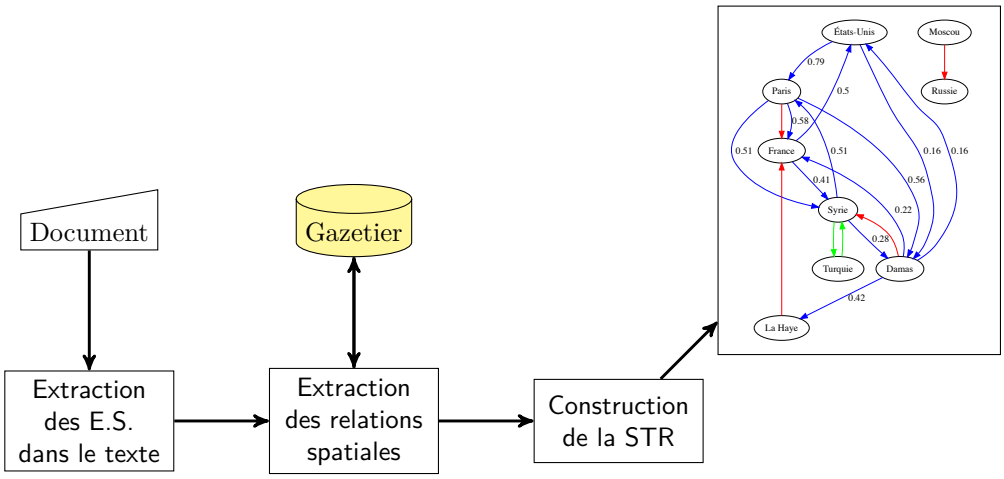

FIGURE *1. Processus de création d'une STR*

dédié. Puis, à partir des différentes entités spatiales et des informations contenues dans le texte et dans le gazetier, on extrait les liens permettant de générer le graphe STR. Enfin, pour mesurer la similarité entre les STR, nous adoptons des mesures de *Graph Matching* (Riesen *et al.*, 2010). Ce processus est détaillé dans les sous-sections suivantes.

3.1. Définitions préliminaires

Une **entité spatiale** est une entité qui peut-être localisée dans l'espace (Casati, Varzi, 1997). L'ensemble des entités spatiales est divisé en deux catégories : entités spatiales nommées (par ex. Madagascar, Syrie, ...) et les entités spatiales abstraites (par ex. ville, route, ...).

La **configuration spatiale** définit la disposition des différents entités spatiales dans un environnement défini.

Un **QCN**, ou Qualitative Constraint Network, est une représentation sous forme de graphe G d'une configuration spatiale d'une liste d'entités I utilisant un ou plusieurs calculs qualitatifs C_i à travers une interprétation qualitative Q_C. Il est défini par $G = (V, E)$ où V correspond à l'ensemble d'entité I et E, les différentes associations possible dans I à l'aide des calculs qualitatifs dans C.

Un **calcul qualitatif** C_i associe deux entités dans une configuration spatiale. Il existe différents types de calcul : *topologie* (RCC-8), *direction* (points cardinaux), *distance*, etc.

Soit un ensemble de calculs qualitatifs C et un ensemble d'entité I, l'**interprétation qualitative** est l'opération Q_C qui relie les entités dans I en utilisant chaque calcul qualitatif C_i de C.

La Figure 2 présente un exemple de QCN utilisant la direction (N, E, S, W) comme calcul qualitatif.

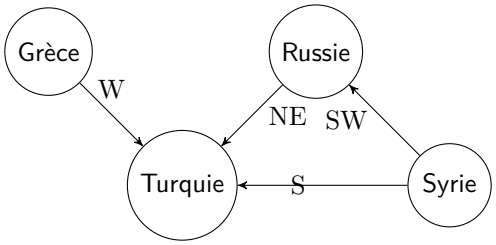

FIGURE *2. Exemple d'un QCN (SW = South West, NE= North East, ...)*

3.2. Création d'une STR

Nous définissons une STR comme un graphe G qui possède un ensemble $V = \{es_1, es_2, ...es_n\}$ de nœuds (i.e. entités spatiales) et un ensemble d'arrêtes $E = \{E_{inc} \cup E_{adj} \cup E_{occ}\}$ divisé selon les différentes relations existantes (inclusion, adjacence, co-occurrence). Alors, une STR est définie par un graphe G :

$$G = (V = \{es_1, es_2, ...es_n\}, E = E_{inc} \cup E_{adj} \cup E_{occ})$$

Le processus de création d'une STR à partir de données textuelles se divise en deux phases : l'extraction des entités spatiales puis des relations associées pour créer le graphe correspondant.

3.2.1. Extraction des entités spatiales

Pour extraire les entités spatiales nommées, il existe plusieurs méthodes dans la littérature, dont la plupart utilisent une ressource externe appelée *gazetier*. Un gazetier est constitué d'un ensemble de données structurées liées à des noms de lieux (ou *toponymes*). Il existe plusieurs gazetiers, dont le plus connu est Geonames [1]. Cependant, pour la majorité des gazetiers, certaines informations sont manquantes et/ou incomplètes. Pour cause, les besoins définis lors de la création de ceux-ci. Le gazetier de Getty [2] a été créé dans le but de

1. http://www.geonames.org
2. http://www.getty.edu/research/tools/vocabularies/tgn/index.html

cataloguer des oeuvres artistiques. Par conséquent, il est normal que les coordonnées des frontières soient absentes. Dans l'objectif de faciliter l'accès aux différentes données nécessaires à la création des STR, mais aussi de proposer un jeu de données simple comportant des données précises, nous avons construit un gazetier appelée Geodict [3] (Fize, Shrivastava, 2017). Au moment de l'écriture de cet article, il contient 4.1 millions d'entrées dont environ 170 000 sont associées aux coordonnées de leur frontière. Les données sont extraites à partir de différentes sources : Wikidata [4], Geonames, OpenStreetMap [5].

La méthode d'identification d'entités spatiales (*geotagging*) dans les textes que nous proposons se déroulent alors en deux étapes :

1. Identification des toponymes ;
2. Association de chaque toponyme à une entité spatiale.

1re étape : Identification des toponymes

Pour identifier les toponymes, nous utilisons un NER ou Named Entity Recognizer. Nous pouvons citer plusieurs NERs, notamment StanfordNER (Finkel *et al.*, 2005), OpenNLP [6] et Polyglot (Al-Rfou *et al.*, 2015). Toujours dans le cadre de la mise en correspondance de données hétérogènes, nous avons choisi Polyglot car il peut être utilisé avec 40 langues ce qui permet d'appliquer notre méthode sur des corpus multilingues. Nous proposons de valider les toponymes retournés par le NER selon leur existence dans le gazetier Geodict.

2e étape : Association de chaque toponyme à une entité spatiale

Il s'agit d'associer chaque toponyme à une entité spatiale présente dans le gazetier. Cependant, il existe, pour certains toponymes, plusieurs entités spatiales possibles *e.g. Paris, France* $\neq$ *Paris, Las Vegas*. Il existe différentes techniques de désambiguïsation (Moncla *et al.*, 2014). Une première solution est d'associer, à chacun de ces toponymes, l'entité spatiale la plus couramment utilisée (Paris $\rightarrow$ Paris, France). Une deuxième solution, plus sophistiquée, est de prendre en compte le contexte géographique du texte. Nous avons choisi la première solution, car les ambiguïtés telles que Paris, France $\neq$ Paris, Las Vegas sont rares dans les corpus sur lesquels nous travaillons. Pour cela, nous associons à chaque entité spatiale la valeur retournée par le calcul du PageRank (Page *et al.*, 1999) sur différents corpus.

3.2.2. Extraction des liens entre les entités spatiales

Après avoir identifié les entités spatiales dans le texte, nous extrayons les différentes relations existantes entre elles. Nous avons choisi trois types de relations : co-occurrence, adjacence (France$\Leftrightarrow$Belgique), inclusion (Coutances $\rightarrow$

3. Disponible ici : http://dx.doi.org/10.18167/DVN1/MWQQOQ
4. https://wikidata.org
5. http://openstreetmap.org
6. https://opennlp.apache.org/index.html

France). Pour extraire les relations d'adjacence et d'inclusion, nous utilisons les informations associées à chaque entité spatiale dans le gazetier. Puis, si deux entités spatiales dans le texte considéré apparaissent l'une à la suite de l'autre, alors elles sont associées avec une relation de co-occurrence.

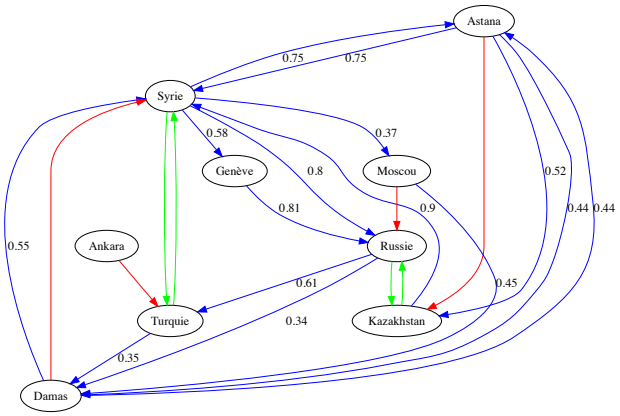

FIGURE *3. Une STR générée à partir d'un article de presse traitant du conseil d'Astana dans la cadre de la guerre civile en Syrie (Légende : bleu = co-occurrence; rouge = inclusion; vert = adjacence)*

3.3. Comment mesurer la similarité entre STR ?

À partir des STRs ainsi construites, nous adoptons des méthodes de comparaison de graphes, connues dans la littérature par *Graph Matching*, pour mettre en correspondance les documents associés.

Dans (Riesen *et al.*, 2010), le **graph matching** est défini comme le processus qui évalue la similarité de deux graphes. Il est possible de distinguer :

– *Exact Graph Matching* : qui retourne une valeur égale à 1 ou 0 (identique ou non) ;

– *Inexact ou Error-Tolerant Graph Matching* : qui retourne une valeur réelle entre 0 et 1.

Dans les mesures proposées dans l'état de l'art, (Bunke, Shearer, 1998 ; Wallis *et al.*, 2001) proposent deux mesures simples (voir Equation 1) et intuitives utilisant le *mcs* ou maximum common subgraph.

$$MCS(g_1, g_2) = \frac{|mcs(g_1, g_2)|}{max(g_1, g_2)} \qquad WGU(g_1, g_2) = \frac{|mcs(g_1, g_2)|}{|g_1| + |g_2| - |mcs(g_1, g_2)|} \tag{1}$$

D'autres mesures existent telles que la GED (Riesen, Bunke, 2009) ou Graph Edit Distance, qui calcule le coût de la transformation d'un graphe A vers un graphe B. Malheureusement, cette mesure implique des temps de calculs plus élevés à mesure que les graphes comparés grandissent et se densifient.

Dans le cadre de nos travaux, nous proposons deux mesures de similarité. La première est fondée sur la combinaison du calcul de l'indice de Jaccard (Levandowsky, Winter, 1971) sur l'ensemble des nœuds et sur l'ensemble des arêtes des deux graphes comparés.

$$sim_{jaccard}(g_1, g_2) = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|} \times \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|} \tag{2}$$

Pour la deuxième mesure proposée, nous utilisons les travaux de (Grover, Leskovec, 2016) qui proposent une représentation vectorielle de la topologie présente autour de chaque nœud d'un graphe. Cette approche est très intéressante car :

– La représentation de la topologie est diminuée à une seule variable : les nœuds. Ceci facilite grandement la comparaison de graphes volumineux.

– Et contrairement à la similarité MCS et WGU, les arêtes ainsi que leur valeur associée sont prises en compte dans la construction des vecteurs relatifs à chacun des nœuds.

Nous proposons ainsi une mesure de similarité, nommée N2V, sur les graphes. On définit N2V se définit par :

$$N2V(G_1, G_2) = \frac{|V_1 \cap V_2|}{|V_1 \cup V_2|} \times (\frac{1}{|D|} \times (\sum_{rel_i \in D} \frac{1}{|N_c|} \times \sum_{i \in N_c} sim_{cos}(n2v(G_{1ij}), n2v(G_{2ij})))) \tag{3}$$

avec

– V_i, ensemble de nœuds du graphe i ;

– D, ensemble des relations présentes dans les graphes ;

– N_c, nœuds en commun des deux graphes ;

– $n2v(G_{1ij})$, représentation vectorielle du nœud j dans le graphe 1 décrit par la relation i ;

– $sim_{cos}(v_1, v_2)$, similarité *cosinus* entre les vecteurs v_1 et v_2.

4. Expérimentations

Les expérimentations sont réalisées sur 3 corpus distincts (voir la synthèse Table 1).

1) CAGN, un corpus généré à partir du flux d'actualité de Google News [7] sur différents thèmes populaires lors de l'extraction [8]: Trump, François Fillon, Alep ;

2) CJG, un corpus composé d'articles de presse traitant de sujets géopolitiques. Ces documents faisant intervenir en moyenne plus d'entités spatiales, nous pensons qu'il est intéressant de tester notre structure sur ceux-ci ;

3) EPI, un corpus composé des résumés de publications scientifiques divisées selon plusieurs épidémies animales : la peste porcine africaine, la fièvre catarrhale et le virus de Schmallenberg.

Corpus	**Type de document**	**Nb. de doc**	**Thématique(s)**	**Langue**
CAGN	article de presse	198	Alep, Trump, François Fillon	FR
CJG	article de presse	27	Géo-politique	FR
EPI	abstract	286	Epidémiologie animale	EN

TABLE *1. Corpus utilisés*

Les évaluations sont réalisées avec le calcul de la *precison at n* (Craswell, 2009) ou P@n. P@n est la proportion des n documents les plus similaires dont la mise en correspondance est pertinente.

Différentes représentations des documents sont évaluées :

– **BOW-SE** : Les documents sont représentés à l'aide d'un bag-of-words avec un vocabulaire composé uniquement des toponymes identifiés dans le corpus concerné. La pondération utilisée est TF-IDF.

– **QCN** : Les documents sont transformés en QCN. Les relations utilisées sont l'adjacence et l'inclusion.

– **STR** : Les documents sont transformés en STR. Les relations utilisées sont l'adjacence, l'inclusion et la co-occurence.

L'évaluation est divisée en deux parties. Dans une première partie, nous confrontons une approche classique BOW-SE à approche QCN. Puis, dans une deuxième partie, nous évaluons la structure STR pour ensuite discuter des résultats comparés.

7. https://news.google.fr/
8. Au cours du premier semestre 2017

4.1. Résultats

4.1.1. Apport des QCN: BOW-SE vs QCN

Corpus	**QCN (MCS)**		**QCN (Jaccard)**		**BOW-SE**	
	P@1	P@4	P@1	P@4	P@1	P@4
CAGN	0,77	0,7	0,43	0,39	0,17	0,16
CJG	0,88	0,68	0,51	0,39	0,11	0,10
EPI	0,56	0,50	0,25	0,29	0,11	0,21
Score Moyen	**0,74**	0,63	**0,40**	0,36	0,13	**0,16**

TABLE *2. Scores obtenus avec les modèles QCN et BOW-SE*

4.1.2. STR

La Table 3 présente les statistiques de comportement de la STR sur les différents corpus.

Corpus	**Taille Moy. Doc. (nb. de mots)**	**Taille Moy. STR**	**% de doc. avec des ES**
CAGN	688.82	4.51	0.83
CJG	643.85	4.33	0.88
EPI	335.05	2.36	0.83

TABLE *3. Données générales sur les différents corpus utilisés*

Corpus	**MCS**		**WGU**		**JACCARD**		**N2V**	
	P@1	P@4	P@1	P@4	P@1	P@4	P@1	P@4
CAGN	0,77	0,71	0,75	0,7	0,49	0,44	0,18	0,29
CJG	0,88	0,68	0,88	0,68	0,51	0,39	0,03	0,08
EPI	0,56	0,49	0,56	0,50	0,27	0,29	0,11	0,21

TABLE *4. Scores obtenus avec la STR*

4.2. Discussion

Dans une première partie, nous avons voulu vérifier ce qu'une approche telle que les QCN pouvait apporter face à des modèles classiques. Bien que l'approche BOW-SE soit simpliste, on remarque que les approches de types QCN obtiennent de meilleurs scores.

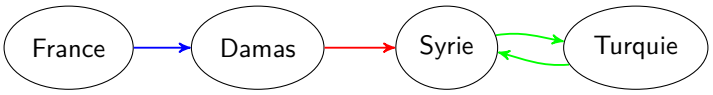

FIGURE *4. Exemple d'une STR moyenne*

Corpus	N2V (STR)	
	P@4	P@20
CAGNews	0,60	0,44
CJG	0,50	-
EPIDEMIO	0,20	0,13

TABLE *5. Évaluation en utilisant uniquement les documents possédant des graphes avec un minimum de 6 nœuds*

L'un des points importants des STR consiste à renforcer ces structures pour qu'elles caractérisent la configuration spatiale mais aussi des relations spatiales propres aux textes : l'enchainement des entités spatiales (relations de co-occurrences) et par la suite des déplacements. À partir des résultats Table 4, on observe que les mesures simples obtiennent de meilleurs performances que les méthodes complexes sur la totalité des corpus

Une des raisons principales se trouve être la taille moyenne des graphes. En effet, les graphes possèdent en moyenne 4 nœuds (voir Figure 4), excepté pour le corpus EPI pour lequel les graphes possèdent en moyenne deux nœuds (voir Table 3). Ceci a pour effet de diminuer fortement les chances que deux graphes possèdent des entités spatiales communes (resp. des relations communes).

Toutefois, il est possible d'avoir des documents qui génèrent des graphes comme celui illustré dans la Figure 5, où l'on observe une richesse avec une structure plus complexe. En effet, le texte, à partir duquel a été générée la STR de la Figure 4, revient sur la preuve apportée par la France sur l'utilisation du gaz sarin par le gouvernement syrien.

Par conséquent, nous avons évalué les différentes mesures uniquement sur des documents associés à des graphes ayant un nombre de nœuds supérieur à la moyenne et plus spécifiquement supérieur ou égal à 6. Les mesures telles que N2V donnent alors de meilleurs scores (voir Table 5). En effet, plus la topologie d'un graphe est complexe, plus ce type de mesure devient pertinent. Par ailleurs, la dimension thématique reste cruciale pour mettre en relation des documents, ce qui explique le bon comportement de l'approche bag-of-words. Ceci nous encourage à combiner notre représentation spatiale STR aux informations thématiques issues des textes dans nos futurs travaux.

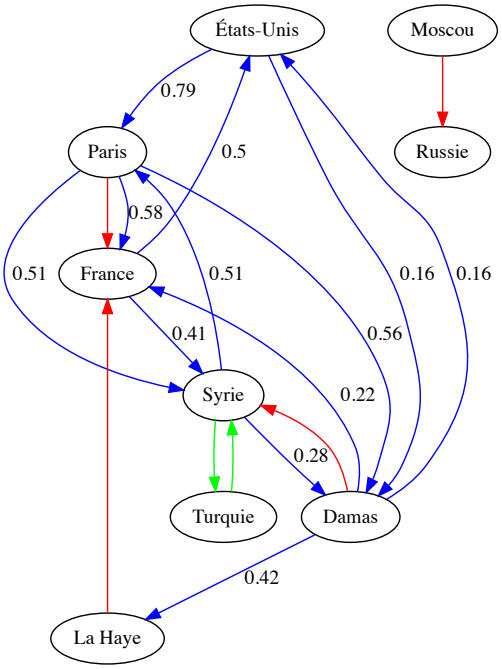

FIGURE 5. *Une STR générée à partir d'un article sur la preuve apportée par la France concernant l'utilisation du gaz sarin par le gouvernement syrien*

5. Conclusion

Dans cet article, nous avons proposé une représentation de la spatialité des données textuelles nommée STR, basée sur une structure de graphe inspirée des QCNs. Dans le cadre de la mise en correspondance de données textuelles, la STR, à travers différentes mesures de similarité, offre des résultats satisfaisants bien que son utilisation reste sensible à la taille et à la densité des graphes générés. Parmi les perspectives envisagées, nous souhaitons implémenter une politique de désambiguïsation plus sophistiquée ainsi qu'une évaluation des comportements de la STR plus approfondie. Nous souhaitons aussi combiner des informations thématiques issues des textes aux STR pour améliorer la tache de mise en relation de documents.

Bibliographie

Al-Rfou R., Kulkarni V., Perozzi B., Skiena S. (2015, April). Polyglot-NER: Massive multilingual named entity recognition. *Proceedings of the 2015 SIAM Internatio-*

nal Conference on Data Mining, Vancouver, British Columbia, Canada, April 30 - May 2, 2015.

Al-Salman R., Dylla F., Fogliaroni P. (2012). Matching geo-spatial information by qualitative spatial relations. *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD '12*, p. 38.

Amitay E., Har'El N., Sivan R., Soffer A. (2004). Web-a-where: geotagging web content. In *Proceedings of the 27th annual international acm sigir conference on research and development in information retrieval*, p. 273–280.

Belouaer L., Brosset D., Claramunt C. (2016). From verbal route descriptions to sketch maps in natural environments. In *Proceedings of the 24th acm sigspatial international conference on advances in geographic information systems*, p. 13.

Blei D. M., Edu B., Ng A. Y., Edu A., Jordan M. I., Edu J. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, vol. 3, p. 993–1022.

Bunke H., Shearer K. (1998). A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters*, vol. 19, n° 3-4, p. 255–259.

Casati R., Varzi A. C. (1997). Spatial entities. In *Spatial and temporal reasoning*, p. 73–96. Springer.

Craswell N. (2009). Precision at n. In L. LIU, M. T. ÖZSU (Eds.), *Encyclopedia of database systems*, p. 2127–2128. Boston, MA, Springer US.

Finkel J. R., Grenager T., Manning C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, p. 363–370.

Fize J., Shrivastava G. (2017). Geodict: an integrated gazetteer. In *Workshop on language, ontology, terminology and knowledge structures, held at the 12th international conference on computational semantics (iwcs 2017)*.

Gaio M., Sallaberry C., Van Nguyen T. (2012). Typage de noms toponymiques à des fins d'indexation géographique. *Traitement Automatique des Langues*, vol. 53, n° 2, p. 143–176.

Grover A., Leskovec J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*.

Le Q. V., Mikolov T. (2014). Distributed representations of sentences and documents. In *Icml*, vol. 14, p. 1188–1196.

Levandowsky M., Winter D. (1971). Distance between sets. *Nature*, vol. 234, n° 5323, p. 34–35.

Ligozat G. (2013). *Qualitative spatial and temporal reasoning*. John Wiley & Sons.

Lin D. *et al.* (1998). An information-theoretic definition of similarity. In *Icml*, vol. 98, p. 296–304.

Moncla L., Renteria-Agualimpia W., Nogueras-Iso J., Gaio M. (2014, novembre). Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed

hiking descriptions corpus. In ACM (Ed.), *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2014)*. Dallas, Texas, United States. Consulté sur https://hal.archives-ouvertes.fr/hal-01069625

Page L., Brin S., Motwani R., Winograd T. (1999). *The pagerank citation ranking: Bringing order to the web.*. Rapport technique. Stanford InfoLab.

Prudhomme C., Homburg T., Jean-Jacques P., Boochs F., Roxin A., Cruz C. (2017). Automatic integration of spatial data into the semantic web. In *Webist 2017*.

Riesen K., Bunke H. (2009). Approximate graph edit distance computation by means of bipartite graph matching. *Image and Vision computing*, vol. 27, n° 7, p. 950–959.

Riesen K., Jiang X., Bunke H. (2010). Exact and inexact graph matching: Methodology and applications. In *Managing and mining graph data*, p. 217–247.

Salton, Mcgill M. (1986). Introduction to Modern Information Retrieval.

Wallgrün J. O., Wolter D., Richter K.-F. (2010). Qualitative matching of spatial information. In *Proceedings of the 18th sigspatial international conference on advances in geographic information systems - gis '10*, p. 300. Consulté sur http://doi.acm.org/10.1145/1869790.1869833{\%}5Cnhttp://portal.acm.org/citation.cfm?doid=1869790.1869833

Wallis W. D., Shoubridge P., Kraetzl M., Ray D. (2001). Graph distances using graph union. *Pattern Recognition Letters*, vol. 22, n° 6/7, p. 701–704.

Wing B. P., Baldridge J. (2011). Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, p. 955–964.

Woodru A. G., Plaunt C. (1994). Gipsy: Georeferenced information processing system,". *Journal of the American Society for Information Science*, vol. 45, n° 9, p. 645–655.

Ye X., Shen H., Ma X., Bunescu R., Liu C. (2016). From word embeddings to document similarities for improved information retrieval in software engineering. In *Proceedings of the 38th international conference on software engineering*, p. 404–415.