Appariement automatique de données hétérogènes: textes, traces GPS et ressources géographiques

Amine Medad, Mauro Gaio, Sébastien Mustiere

Appariement automatique de données hétérogènes: textes, traces GPS et ressources géographiques

Amine Medad[1], **Mauro Gaio**[1], **Sébastien Mustière**[2]

1. *LIUPPA, Université de Pau et des Pays de l'Adour, 64013 Pau Cedex, France*
   im.medad@univ-pau.fr,mauro.gaio@univ-pau.fr
2. *UPE, IGN ENSG, LASTIG COGIT*
   Sebastien.Mustiere@ensg.eu

RÉSUMÉ. *Les travaux que nous présentons dans cet article sont réalisés dans le cadre du projet ANR Choucas. Nous proposons une approche pour l'appariement automatique de traces, de textes de description de randonnées et de ressources géographiques (gazetiers et bases de données géographiques). L'objectif de cet article est d'exposer les premiers éléments d'une méthodologie d'appariement automatique dont les trois étapes sont : l'annotation des traces GPS, l'identification des entités nommées spatiales dans des textes décrivant des itinéraires de randonnées, et la mise en correspondance des toponymes.*

ABSTRACT. *The work we present in this article is carried out within the Choucas ANR project. We propose an approach for the automatic matching of traces, hiking description texts and geographical resources (gazetteers et geographical data bases). The main goal of this article is to expose the first methodological elements for the automatic matching of hiking descriptions and GPS traces. To achieve this goal three steps are needed: GPS traces annotation, spatial named entities identification and toponyms matching.*

MOTS-CLÉS : *Extraction d'information, traitement automatique du langage, données réparties et hétérogènes, appariement*

KEYWORDS: *Information retrieval, Natural language processing, heterogeneous and distributed data, matching*

1. Introduction

Grâce aux plateformes de partage d'informations de randonnées (VisoRando, Camp2camp, IGNRando, ... etc.), nous avons de plus en plus accès pour chacune d'entre elles à une description textuelle de l'itinéraire à suivre souvent accompagnée d'une trace GPS. Les auteurs de ces textes utilisent dans leurs descriptions des points de repères faciles à identifier mais qui compte tenu de leurs caractéristiques en termes de taille, de notoriété ou alors de la façon de les nommer, peuvent ne pas être présents dans les ressources géographiques.

Nos travaux de thèse ont pour but de répondre à la problématique posée par le projet ANR Choucas et plus particulièrement la structuration de données issues de sources textuelles hétérogènes, qui est présentée et expliquée dans la section 2 de l'article (Olteanu-Raimond *et al.*, 2017). Pour ce faire nous cherchons à mettre en correspondance des données hétérogènes : textes narratifs, traces GPS et ressources géographiques (gazetiers, BD géographiques et LOD « données ouvertes liées »), la solution à cette problématique implique la nécessite de répondre à plusieurs verrous scientifiques d'ordre :

1. Cognitif et linguistique : la complexité de la description via le langage naturel de concepts spatiaux décrivant des situations spatiales statiques (ex : description de la position d'une victime) ou dynamiques (ex : description d'itinéraire par les randonneurs) est un problème qui nécessite d'établir une passerelle entre les domaines de la linguistique, des sciences cognitifs, et de l'informatique (traitement automatique du langage naturel). La réponse à cette problématique nécessite la conception de motifs pour identifier les indices linguistiques de relations spatiales dans des textes de randonnées.

2. Raisonnement spatial : peu de travaux s'intéressent à l'appariement de données issues d'un texte (l'itinéraire décrit) avec des données spatiales (la trace GPS). La randonnée peut être vue comme une suite de relations spatiales entre points de repères impliqués dans des actions de mouvement, cela implique d'une part, une modélisation opératoire des actions de mouvement, des relations et des entités extraites du texte et un choix non triviale du modèle de définition des relations spatiales le plus adapté, et d'autre part, la mise en place des mécanismes d'analyse spatiale permettant de raisonner sur des informations souvent à caractère implicite ou incomplet.

Dans cet article nous décrivons les premières étapes d'une méthode permettant de faire l'appariement entre les informations contenues dans les textes et les informations issues des traces GPS associées. L'appariement permettra de faire correspondre à chaque point de repère évoqué dans le texte une géolocalisation ou un ensemble de géolocalisations obtenues à partir de la trace GPS associée. En d'autres termes pour chaque point de repère : soit la localisation la plus plausible obtenue à partir des ressources, soit une localisation approximative pour ceux qui ne sont initialement pas répertoriés dans les ressources géographiques. Ainsi, ce travail permettra à terme de fournir des données de

randonnées enrichies (trace annotée avec des éléments textuels de localisation issu des textes de randonnées) et d'enrichir des bases de données de points de repères en montagne.

L'article est organisé comme suit : la partie 2 présente les données (textes, traces GPS et ressources géographiques) sur lesquelles va porter notre travail. Dans la partie 3 nous expliquons comment nous comptons faire l'appariement entre ces différentes données. La partie 4 est dédiée à une expérimentation sur une première tentative d'appariement. Dans la partie 5 nous concluons et proposons des perspectives.

2. Description des données manipulées

2.1. Les textes

Les textes décrivant des randonnées sont de type narratif. Ils se présentent sous la forme d'un discours comportant des descriptions spatiales. Ces descriptions sont essentiellement composées d'indications sur le chemin à prendre pour atteindre les différentes étapes constitutives de l'itinéraire à parcourir. Les indications n'apparaissent pas forcement dans l'ordre dans lesquelles les étapes seront réalisées. L'exemple ci-après montre un extrait de texte décrivant une randonnée :

" À partir du port de Carantec, traverser la voie submersible. L'île étant tout en longueur, il suffit de suivre le chemin principal jusqu'au bout de l'île puis de revenir sur ses pas. L'île abrite quelques maisons mais surtout une superbe petite chapelle visible de loin car perchée sur une butte ..."

L'indication donnée dans le texte est fréquemment formulée sous la forme d'une phrase contenant un verbe de mouvement, une ou plusieurs relations spatiales et une ou plusieurs entités nommées spatiales étendues (ENSE) [1] (Gaio, Moncla, 2017), le sujet étant généralement le randonneur. Dans la section 3 nous présentons comment nous comptons extraire les ENSE à partir de ces textes.

2.2. Les traces GPS

Les plateformes de partage d'informations de randonneurs associent souvent une trace GPS à chaque description de randonnée. Les traces GPS peuvent être

1. ENSE est une entité construite à partir d'un nom (propre ou commun) et d'un ou plusieurs concepts relatifs à l'expression de la localisation dans la langue, plusieurs niveaux d'encapsulation ont été définis chaque niveau (n) est encapsulé dans le niveau précèdent ($n-1$), ex : Paris est une ENSE de niveau 0, Mairie de Paris est une ENSE de niveau 1. Dans notre contexte de randonnée, les points de repères, lieux et point de passages sont considérés comme des ENSE

interprétées comme une autre représentation des déplacements décrits dans les textes en langage naturel. Elles sont souvent disponibles sous la forme de fichiers au format GPX [2] contenant obligatoirement une collection de coordonnées utilisables sous la forme de points de cheminement, trace ou itinéraire et de manière optionnelle d'autres informations.

Les traces GPS disponibles sur les plateformes contiennent des coordonnées 2D (Latitude, Longitude), parfois 3D (Latitude, Longitude, Altitude), avec ou sans horodatage (timestamp), et en général aucune autre information. Il est également important de souligner que les traces sur lesquelles nous travaillons sont en général des traces corrigées, où les positions aberrantes (Buard *et al.*, 2015) ont été supprimées au préalable.

2.3. Les ressources géographiques

Les ressources géographiques que nous utilisons se présentent sous la forme de gazetiers (Geonames, OpenStreetMap, GoogleMaps), de base de données géographiques (BD Topo, Base adresse nationale), et de LOD (DBPedia). Les informations dans ces ressources sont parfois redondantes, d'autres fois contradictoires, mais elles sont souvent complémentaires. Ces ressources externes sont accessibles via des services Web permettant de les interroger grâce à leurs API.

3. Appariement automatique des données

Pour faire l'appariement de ces données hétérogènes notre démarche consiste à utiliser les toponymes comme éléments d'appariement. L'appariement est divisé en 3 étapes : annotation des traces GPS, identification des différentes ENSE dans les textes de randonnées, mise en correspondance des informations extraites.

3.1. Annotation de la trace GPS

La technique dite du géocodage inverse permet, à partir de gazetiers, d'attribuer une adresse ou un toponyme à des coordonnées géographiques (points de la trace). Les ressources géographiques interrogées, via une paramétrisation adaptée (exemple dans le cas de l'API OSM : Zoom = 18 [3], Namedetails = 1 [4]), renvoient le toponyme le plus proche selon certains critères.

---

2. GPX (eXchange Format) est un format de fichiers qui permet l'échange de coordonnées entre les dispositifs GPS ou des applications de visualisation de traces GPS. Les fichiers GPX se présentent sous la forme de fichiers XML avec une syntaxe contrôlée par un modèle.

3. Paramètre qui permet de définir le niveau de détail requis, la valeur 0 correspond au niveau pays, tandis que la valeur 18 correspond au niveau maison/bâtiment.

4. Permet d'inclure une liste de noms alternatifs dans les résultats. Celle-ci peut inclure des variantes linguistiques, et des abréviations.

Dans notre contexte pour faire du géocodage inverse il n'existe pas, à notre connaissance, de ressource géographique faisant référence. Pour l'annotation de nos traces nous avons utilisé quatre ressources complémentaires : Geonames [5], OpenStreetMap (OSM) [6], GoogleMaps [7], Base Adresse Nationale [8]. Contrairement aux gazetiers (Geonames, OSM, Google) qui sont internationaux, la Base Adresse Nationale est comme son nom l'indique, une base de données des adresses postales françaises.

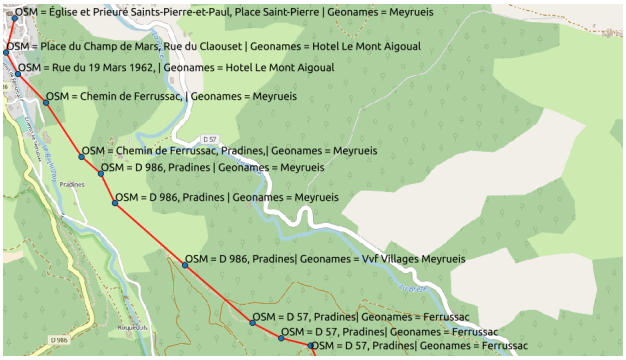

FIGURE *1. Illustration d'une trace GPS annotée*

La figure 1 montre le résultat de l'annotation d'une trace GPS. Chaque point est annoté avec le résultat de son géocodage inverse. Par souci de lisibilité nous n'avons affiché que le résultat issu des ressources OSM et Geonames.

3.2. Geoparsing des ENSE à partir des textes de randonnées

À partir des textes nous cherchons à extraire les lieux (toponymes) et points de passage (ex : Place AlberHuille, La Forge, col de la Madeleine), appelés entités nommées spatiales étendues (ENSE) (Gaio, Moncla, 2017). Le *geoparsing* est la tâche qui permet d'extraire des ENSE dans des textes. Dans la littérature on retrouve plusieurs outils de *geoparsing* (Geoparseurs). Pour les langues romanes (français, espagnol, italien), (Moncla *et al.*, 2016) proposent un système appelé PERDIDO qui permet de reconstruire automatiquement des itinéraires à partir de textes décrivant des randonnées, le système se base sur : une cascade de transducteurs pour la phase d'annotation (*geotagging*), des ressources externes pour la phase de *geocoding*, et une méthode de clustering pour la

5. http://api.geonames.org/findNearbyJSON.
6. http://locationiq.org/v1/reverse.php
7. https://maps.googleapis.com/maps/api/geocode/.
8. https://api-adresse.data.gouv.fr/reverse/.

phase de désambiguïsation. PERDIDO a été conçu et évalué sur des textes de descriptions d'itinéraires en français, c'est principalement pour cette raison que nous avons choisi d'utiliser ce système [9] pour l'extraction des ENSE. Le système PERDIDO retourne un texte où les ENSE, du texte, sont annotées. Quand cela est possible, le système associe également une géolocalisation à chaque ENSE.

3.3. Mise en correspondance des toponymes

Après avoir annoté les traces et extrait les ENSE des textes correspondants, nous passons à la phase d'appariement. Nous présentons ici une première expérimentation d'appariement permettant de faire une évaluation préliminaire du potentiel de l'approche. Nous avons décidé de faire un appariement ponctuel, qui consiste à attribuer à chaque ENSE extraite du texte un ou plusieurs points sur la trace GPS. Nous comparons de manière récursive les ENSE de différents niveaux du texte et les toponymes de la trace annotée, en commençant par l'EN de plus bas niveau (EN_0). Nous choisissons ensuite l'ENSE de niveau le plus haut qui a une correspondance exacte avec un toponyme issu du geocodage inverse de la trace (algorithme 1).

4. Résultats de l'appariement

L'expérimentation préliminaire a été réalisée sur un corpus de 8 descriptions de randonnées issues de VisoRando [10] chacune accompagnée de la trace GPS associée. Le tableau 1 montre pour chaque texte du corpus le nombre d'ENSE dans le texte qui correspondent à des toponymes issus du géocodage inverse des points de la trace associée.

En analysant ces résultats, on remarque que sur 115 ENSE reconnues par PERDIDO, 61 trouvent une correspondance avec les toponymes issus du géocodage inverse de la trace, ce qui fait un rappel de 53%.

Ces scores montrent bien la pertinence de notre approche qui permet de détecter à ce stade la moitié des toponymes. Les échecs de mise en correspondance sont la conséquence du non traitement de deux cas spécifiques d'entités spatiales : les premiers sont les lieux-dits et les microtoponymes absents des ressources géographiques, les seconds sont les ENSE construites uniquement à partir d'un nom commun tel que (tourner à droite de l'église, marcher jusqu'à l'auberge) qui ont une localisation implicite et relative au contexte spatial de la randonnée.

9. http://erig.univ-pau.fr/PERDIDO/
10. https://www.visorando.com/

Algorithme 1 : Algorithme d'appariement

Data : L_{ENSE}, L_{GPS} ;
/* L_{ENSE} : Liste des ENSE extraite du texte avec Perdido */
/* L_{GPS} : La liste des points GPS de la trace. */
Result : L_{ENSE}, L_{GPS} ;

```
1  initialization;
2  i = 0 ;
3  j = 0 ;
4  while (j ≤ nb_ENSE) ∧ (costPath_j ≤ 0) do
5      while j ≤ nb_GPS do
6          match = false ;
7          k = 0 ;
8          /* K : le niveau de l'ENSE */
9          while not(match) || k ≤ nb_ENSE do
10             if (Equal(L_ENSE[i][k], L_GPS[j])) /* Equal : Une méthode
                   permettant de vérifier l'égalité stricte de
                   deux chaînes de caractères */
11             { match = true ;
12             i++;
13             j++ ;
14             }
15             else {k++ ;
16             }
```

TABLE *1. Appariement trace GPS/ textes de randonnées*

Titre de la randonnée	Nombre d'ENSE dans la randonnées	Nombre d'ENSE trouvées dans la trace
Belvédère du mont grêle	11	2
Au pays Risle	6	2
Autour de Cussac	7	7
Fontaine de Vaucluse	16	13
Au Pays de Cunlhat	36	19
Sur les crêtes du Doubs	17	9
Autour de Jongieux	11	5
Le Barroux de Malaucène	11	4

5. Conclusion

Dans cet article, nous avons proposé une première méthode pour l'appariement de textes de randonnées et de traces associées. La méthode d'appariement proposée doit être raffinée pour traiter les différences d'écritures entre les ENSE dans le texte et les toponymes dans la trace. Une amélioration possible

serait d'introduire une mesure de similarité comme proposée par (Nguyen *et al.*, 2013). Nous comptons par la suite passer à un autre niveau d'appariement qui est l'appariement des relations spatiales entre les ENSE extraites du texte, avec des fragments de la trace GPS. À terme notre méthode devra être capable de prendre en compte les ENSE basées sur un nom commun ou celles dont le toponyme n'est pas répertorié dans les ressources géographiques utilisées.

Bibliographie

Buard E., Devogele T., Ducruet C. (2015). Trajectoires d'objets mobiles dans un espace support fixe. *Revue Internationale de Géomatique*, vol. 25, n° 3, p. 331–354.

Gaio M., Moncla L. (2017). Extended named entity recognition using finite-state transducers: An application to place names. In *The ninth international conference on advanced geographic information systems, applications, and services (GEOProcessing 2017)*, p. 15–20. Nice, France, IARIA. (<hal-01492994>)

Moncla L., Gaio M., Nogueras-Iso J., Mustière S. (2016). Reconstruction of itineraries from annotated text with an informed spanning tree algorithm. *International Journal of Geographical Information Science*, vol. 30, n° 6, p. 1137–1160.

Nguyen V. T., Sallaberry C., Gaio M. (2013). Mesure de la similarité entre termes et labels de concepts ontologiques. In *Coria 2013*, p. 415–430. Neufchâtel, Suisse. (<hal-00847528>)

Olteanu-Raimond A.-M., Davoine P.-A., Gaio M., Gouardéres E., Van Damme M.-D., Villanova-Oliver M. *et al.* (2017, novembre). Projet CHOUCAS : Intégration de données hétérogènes et raisonnement spatial pour l'aide à la localisation des victimes en montagne. In *Spatial Analysis and GEOmatics 2017 (Sagéo 20017).* Rouen, France. (<hal-01649156>)