Mapping Heterogeneous Textual Data: a Multidimensional Approach based on Spatiality and Theme

Jacques Fize, Mathieu Roche, Maguelonne Teisseire

Mapping Heterogeneous Textual Data: a Multidimensional Approach based on Spatiality and Theme

Jacques Fize[1,2], Mathieu Roche[1,2], and Maguelonne Teisseire[2]

[1] CIRAD, UMR TETIS, F-34398
firstname.lastname@cirad.fr
[2] TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, IRSTEA - Montpellier, France
maguelonne.teisseire@irstea.fr

Abstract. In this paper, we propose a multidimensional mapping approach for heterogeneous textual data that exploits firstly the spatial dimension and secondly the thematic dimension. Based on the Spatial Textual Representation (STR) as well as the Geodict geographic database, the contribution presented in this paper integrates the thematic dimension of documents. To support our proposal on mapping textual documents, we evaluate the different aspects of the process using two real corpora, including one corpus that is highly heterogeneous.

Keywords: Text Mining · Spatial and Thematic Dimensions · Heterogeneous Data

1 Introduction

Over the last few decades, the improvement of data collection and storage techniques has raised new issues around the Big Data area. Big Data is characterized by three V: Volume, Velocity and Variety. In this work, we focus on the heterogeneous dimension of data (the Variety concept), and more particularly on the mapping of textual data. The heterogeneity of textual data is characterized by the variety of structures (e.g. narrative, data table, etc.), formats (e.g. txt, xls, pdf, etc.), sources (newspapers, social networks, etc.), language or vocabulary used. To establish matching links for such data, the design of new representation models and appropriate similarity measures are required. We thus propose a multidimensional mapping approach for heterogeneous textual data that exploits firstly the spatial dimension and secondly the thematic dimension. For the spatial mapping of textual data, we have already proposed a representation of document: the Spatial Textual Representation (STR) [6]. STR is a graphical representation, composed of connected spatial entities (i.e. vertices) according to their spatial relationships (i.e. edges).

In this paper, we focus on the integration of the thematic dimension of document and present a new contribution on mapping textual documents. Generally,

the thematic dimension of a document is associated with a set of keywords (or phrases). These keywords can be obtained automatically or by using resources built by experts (ontologies, thesaurus, dictionaries). We propose to use different collections of keywords (terminologies) associated with the general themes of our corpora. In the context of the thematic mapping, we integrate terminologies from the study case domains and obtained by text-mining approaches. Finally, we propose a new transformation of the STR based on these terminologies to answer the following questions: What are the themes that connect spatial entities? Can these new relationships improve the mapping of the spatial dimension? To support our contributions, we evaluate the different aspects of the process using two corpora, including one corpus that is highly heterogeneous. The results obtained on these real data show that our proposals that integrate thematic information to the spatial dimension improve the task of linking heterogeneous data.

2 Theme Relationships

2.1 Preliminaries

A spatial entity is located in a defined space and associated with 4 properties:

1. A **toponym** (by language) and several aliases, or other names, *e.g. La Ville Lumière (The Light City)* ≡ *Paris*
2. A **geographic footprint** composed of coordinates *latitude-longitude*, and if available, the boundarys geometry described by a Polygon.
3. A **class** defining the nature of this entity
4. **Related information** such as country membership, neighboring entities, demographics if associated with a population.

To compare documents through the spatial dimension, a common representation is required. To this end, in previous work, we defined the **STR** [6], or **S**patial **T**extual **R**epresentation, a graph structure generated from text features. A STR is composed of two components: spatial entities (i.e., vertices) and spatial relations (i.e., edges). A spatial entity is an entity located in a defined space [4] and associated with spatial information — coordinates, country, boundaries, etc. — collected from a georeferential (or gazetteer). A spatial relation connects two spatial entities, *e.g., adjacency, inclusion, distance*. As geographical index, we adopt GEODICT[3] [7], generated from three sources: *Wikidata*, *OpenStreetMap* and *Geonames*. In [6], we conducted experiments with graph matching algorithms [5] to match spatial features integrated into the STR. We defined three STR transformations (two STR Abstractions that abstract spatial entities in the STR - Uniform Abstraction and Bounded Abstraction and one STR Extension) and evaluate the associated matching quality through four criteria (See Section 2.2). Figure 2 illustrates an STR and its associated extension (adjacency relations in green and inclusion relations in red).

[3] Geodict is available at this address: `http://dx.doi.org/10.18167/DVN1/MWQQOQ`

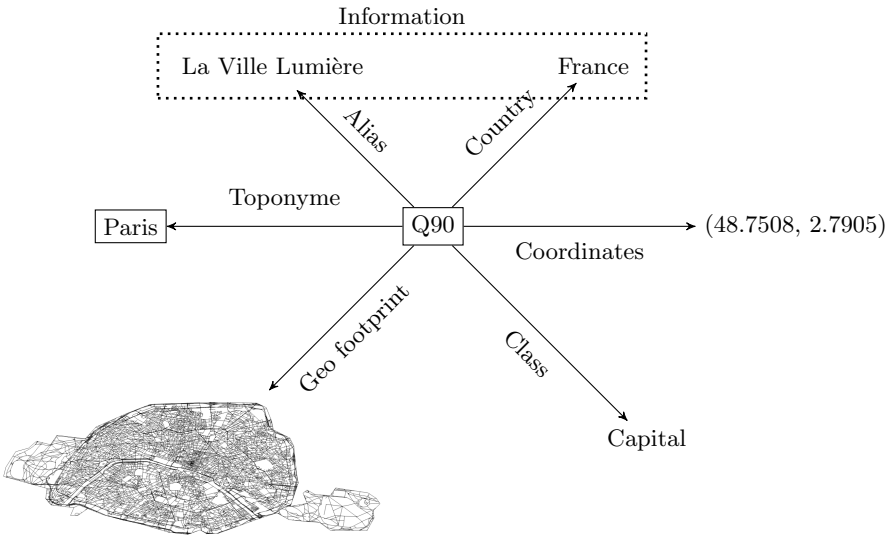

Fig. 1: Information related to the Paris spatial entity

2.2 Thematic entities

The aim is to add new information to the defined STR structure in order to improve the matching process of documents. The hypothesis is that linking thematic concept to spatial entities will add valuable value to the representation. The main issue is to automatically extract thematic entities related to the documents. **A thematic entity** is associated with a phrase [4] as well as a set of variations used according to language or context.

Thematic Integration in the STR structure The thematic entity extraction process is context-sensitive. For each specific corpus, different terminologies from dictionaries, thesauri and ontologies are needed (See Section 3.1). Once the thematic entity extraction is performed, the next step is to rely these entities to the spatial entities. In our work, a thematic relationship exists between a spatial entity and a thematic entity if they belong to the same *window* in the document. This window can be defined in different ways:

1. Pattern based,

   Example 1. [Determinant] [city or region or department] + [Determinant] + [**Spatial Entity**] + [Verb] + [Determinant] + [**Thematic Entity**]
   ⇒
   "*The city of* ***Montpellier*** *retrieves* ***organic waste*** *to produce fertilizer [...]*"

[4] Group of morphemes or words that follow each other with a specific meaning

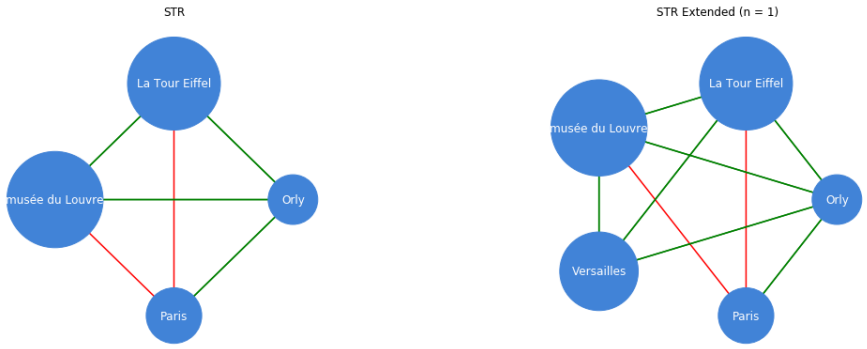

Fig. 2: Example of a STR and its extension

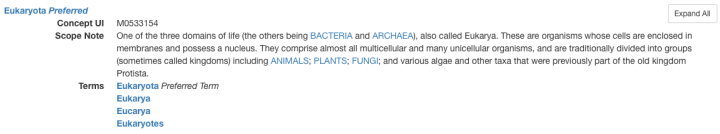

Fig. 3: A thematic entity (Eukaryota) and its variations

2. Sentence based. The sentence is usually the window chosen to extract relation between entities along with features (Part-of-Speech, Dependency, or Word Embedding)
3. Sliding window based.

sliding window of size 5

Example 2. The city of **Cerberus** helps **farmers** in their transition to organic agriculture .

Figure 4 illustrates the enrichment of the STRs presented Figure 2 by adding the corresponding thematic nodes and links (thematic nodes in red and thematic links in blue).

If the corpus is highly heterogeneous, pattern extraction or sentence identification could be a real issue. For this reason, we chose to use the *sliding* window to identify relationship between the spatial entities and thematic entities.

Matching process In [6], the matching quality was evaluated through four criteria:

– Shared Spatial Entities (SSE). The criterion is validated if the two STRs share at least one common spatial entity.
– Close Spatial Entities (CSE). The criterion is validated if one or more spatial entities share a proximity with a different entity in the other STR.

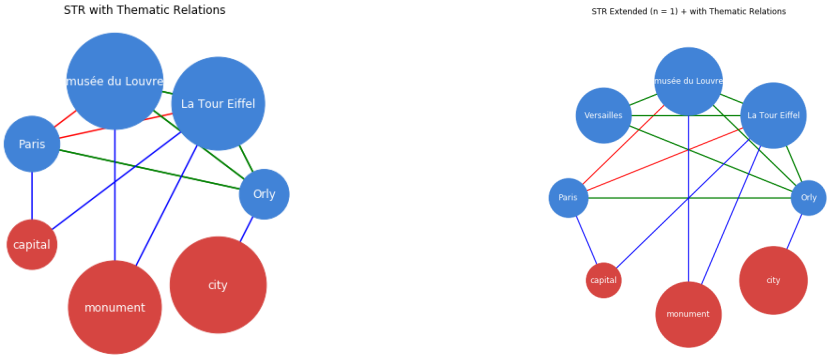

Fig. 4: Example of STRs enriched with thematic nodes (in red) and thematic links (in blue)

- Significant Spatial Coverage (SSC). Dense and significant groups of spatial entities located close to one another may be found in STR.
- Strict Spatial Coverage (SCSC). Distinct groups of spatial entities can be found in STRs.

In this paper, we add two new evaluation criteria based on the feedbacks of experts on the previous results given in [6]:

- Average Distance (AM). It is based on geographical distance average between the space entities belonging to the STRs. The distance is normalized between 0 and 1 then reversed (1 - average distance).
- Percentage of shared spatial entities (PSSP). To measure the similarity between the two sets of spatial entities of two corresponding STRs, we use the Srensen-Dice index (or Dice coefficient) defined as follows:

$$PSSP(STR_i, STR_k) = \frac{2 * |ES_{STR_i} \cap ES_{STR_k}|}{|ES_{STR_i}| + |ES_{STR_k}|} \quad (1)$$

with ES_{STR_i} the set of spatial entities of STR_i.

The evaluation process is then performed through 6 criteria. The best results are the ones maximizing the overall criteria, see Table 1.

3 Experiments

3.1 Datasets and thematic ressources

In this paper, the first corpus is called **PadiWeb** and corresponds to a set of press articles related to animal epidemiology. The second is called **AgroMada**

and corresponds to a set of data produced during a project related to agroecology.

AgroMada corpus. In the last decades, CIRAD has been engaged in developing sustainable agricultural practices in Madagascar. During this period, the CIRAD has produced a significant amount of data including theses, reports, technical manuals, field data, and presentations. Compared to Padi-Web, this corpus is highly heterogeneous. The original corpus is composed of 13 742 documents in different file formats. Based on this corpus, we selected documents corresponding to a specific thematic focus (i.e. agroecology) and a specific location (i.e. Madagascar). The final corpus, AgroMada, is composed of 5552 documents in English and French.
For this corpus, we have selected **4** vocabularies to evaluate the thematic entity extraction:

- INRA, Thesaurus formed from different dictionaries produced by INRA[5],
- Dev.Du., the vocabulary of sustainable development proposed by the French Ministry of Culture [6],
- Biotex BVLAC, vocabulary extracted using the Biotex software [8],
- Biotex + LDA BVLAC The Biotex software combined with LDA (Latent dirichlet allocation) [2].

Padi-Web corpus. Padi-Web [1] is an epidemiology surveillance system implemented by CIRAD (Agricultural Research for Development) in collaboration with INRA (French National Institute for Agricultural Research). Padi-Web produces a classification and extracts information from unofficial sources (Google News) dealing with the emergence of epidemics to remedy delays in the publication of official decrees. To evaluate the volume and accuracy of the extracted information, a gold standard corpus, composed of 500 documents, was built.
Concerning this corpus, we selected **3** vocabularies and terminologie to evaluate the thematic entity extraction:

- BiotexPadi500 → Vocabulary extracted using the Biotex software on the PadiWeb corpus,
- BiotextPadi35k → Vocabulary extracted using Biotex software on a 35000 corpus document extracted using PadiWeb,
- Disease Infect. → Terminologie providing a list of names of infectious animal diseases, as well as their variations.

3.2 Results

As in [6], we selected a panel of graph matching algorithms. We only mention the ones outperforming among the others: As **Structure-based algorithms**, the

[5] https://dicoagroecologie.fr/en/
[6] `http://www.culture.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Enrichissement-de-la-langue-francaise/FranceTerme/Vocabulaire-du-developpement-durable-2015`

Vertex/Edge Overlap [9](VEO), the Maximum Common Subgraph [3](MCS), a measure derived from the Jaccard Index. As **pattern-based algorithms**, the **Bag of Cliques(BOC)** that exploits the connectivity represented by the spatial relations. And we use a text baseline approach, called **BOWSE** that computes the cosine similarity between each document vector, composed of the frequency of toponyms in the document. Table 1 shows the best results of the combination algorithm, type of STR and terminology used.

Corpus	Filter	All Criteria		
		Alg. GM	Type of STR	Terminology
AgroMada	none	MCS	STR	
	format = doc	BOWSE	STR extended	
	format = docx	BagOfCliques	STR	Biotex + LDA BVLAC
	format = html	VertexEdgeOverlap	STR extended	Inra
	format = pdf	PolyIntersect	STR Ext. 2	
	format = xls	VertexEdgeOverlap	STR Ext.1	
	size >= 46	BagOfCliques	STR	Biotex + LDA BVLAC
	size >= 72	BOWSE	STR Ext. 1	
PadiWeb	none	MCS	STR	Disease Infect.
	size >= 4	MCS	STR	Disease Infect
	size >= 7	MCS	STR	Biotex + Padi 35k

Table 1: Best results obtained for the different combinations (STR type, graph-matching algorithm, Vocabulary) on datasets, with or without filters

Integrating relations between spatial entities and thematic entities in the matching process provides good performances. This is particularly true for the PadiWeb corpus which is really specialized. Among the terminologies presented, we observe a large proportion of terminologies derived from the automatic extraction (but cleaned) of syntagms by Biotex.

4 Conclusion

In this paper, we propose a new matching approach for textual document based on the spatial and thematic dimensions. We explore in more details how thematic information could improve the performances. Results are interesting and show disparate impacts according to the context of the data used. Specific corpora seems to have more powerful effect when taking the two dimensions into account. The terminologies used for thematic entity extraction play a crucial role that needs to be deepened in future works.

References

1. Arsevska, E., Roche, M., Falala, S., Lancelot, R., Chavernac, D., Hendrikx, P., Dufour, B.: Monitoring disease outbreak events on the web using text-mining approach and domain expert knowledge. European Language Resources Association (ELRA), Paris, France (may 2016)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (Mar 2003)
3. Bunke, H., Allermann, G.: Inexact graph recognition matching for structural pattern. Pattern Recognition Letters **1**(May), 245–253 (1983). https://doi.org/10.1016/0167-8655(83)90033-8
4. Casati, R., Varzi, A.C.: Spatial entities. In: Spatial and Temporal Reasoning, pp. 73–96. Springer (1997)
5. Fischer, A., Riesen, K., Bunke, H.: Improved quadratic time approximation of graph edit distance by combining Hausdorff matching and greedy assignment. Pattern Recognition Letters **87**, 55–62 (2017). https://doi.org/10.1016/j.patrec.2016.06.014
6. Fize, J., Roche, M., Teisseire, M.: Matching heterogeneous textual data using spatial features. In: 2018 IEEE International Conference on Data Mining Workshops, ICDM Workshops, Singapore, Singapore, November 17-20, 2018. pp. 1389–1396 (2018). https://doi.org/10.1109/ICDMW.2018.00197
7. Fize, J., Shrivastava, G.: Geodict: an integrated gazetteer. Association for Computational Linguistics (2017)
8. Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire, M.: Biomedical term extraction: overview and a new methodology. Inf. Retr. Journal **19**(1-2), 59–99 (2016). https://doi.org/10.1007/s10791-015-9262-2
9. Papadimitriou, P., Dasdan, A., Garcia-Molina, H.: Web graph similarity for anomaly detection. Journal of Internet Services and Applications **1**(1), 19–30 (May 2010). https://doi.org/10.1007/s13174-010-0003-x