Associating spatial patterns to text-units for summarizing geographic information

Julien Lesbegueries
LIUPPA - Université des Pays de l'Adour
Avenue de l'université
BP 576 – 64012 PAU Cedex
00 33 5 59 40 75 70
julien.lesbegueries@univ-pau.fr

Christian Sallaberry
LIUPPA - Université des Pays de l'Adour
Avenue de l'université
BP 576 – 64012 PAU Cedex
00 33 5 59 40 75 70
christian.sallaberry@univ-pau.fr

Mauro Gaio
LIUPPA - Université des Pays de l'Adour
Avenue de l'université
BP 576 – 64012 PAU Cedex
00 33 5 59 40 75 70
mauro.gaio@univ-pau.fr

ABSTRACT
Retrieving data based not only on key words is a challenge. We worked on semi-structured data (cultural heritage corpora). Our project aimed at getting the most relevant text-units of documents (sets of sentences, paragraphs, sections, etc.) according to a spatial query. This paper proposes a method to build summarized spatial indexes for text-units based on spatial patterns. This approach adds semantic interpretation to classical indexing methods.

Categories and Subject Descriptors
H.3.1 Content Analyzing and Indexing: *linguistic processing, indexing methods*
H.3.7 Digital Libraries

General Terms
Management, Experimentation

Keywords
Spatial Information Extraction, Spatial Information Summarization, Spatial Model, Digital Libraries, Semi Structured Data, Cultural Heritage

1. INTRODUCTION
Spatial information extraction (IE), retrieval (IR) and visualization (IV) are the main goal of the "Virtual Itineraries in the Pyrenees[1]" (PIV) prototype [5]. PIV corpora are composed of a specific digital documents library strongly related to the Pyrenean cultural heritage. PIV proposes a spatial model (detailed in [5]) based on the linguistic hypothesis that a spatial feature[2] (SF) is defined from landmarks (named entities) [6] and spatial relationships. This model supports absolute and relative SFs. Named SFs such as "Biarritz district" are well-known named places. We call them Absolute SFs (A_SF). Complex SFs such as "Biarritz vicinity" or "South of Biarritz district" need some linguistic and spatial reasoning processes. Such features are called Relative SFs (R_SF). We associate each R_SF to one or more spatial relationships (adjacency, inclusion, distance, orientation), derived from the qualitative spatial reasoning area, for a recursive definition [5,8]. Therefore, PIV indexes manage spatial core model instances: A_SFs and R_SFs are described by their names, types spatial relationships and geo-located footprints.

A_SFs and R_SFs describe spatial information at sentence-chunk level. In this paper we aim at analysing semantics content of sets of SFs in order to build summarized spatial indexes. As we consider book-structure granularity, spatial summarization stage consists in processing all the paragraphs of one hierarchical level. This process associates spatial patterns to text-units. We propose three patterns: view-point, itinerary and area-comparison. From heuristics and linguists' works concerning texts summarization [7], we established a grid of criteria allowing to attach a SFs set to a pattern with a particular degree of reliability.

The paper is organised as follows. Section 2 presents spatial information summarization basis. In section 3, we describe our SFs summarized indexing. Finally, we illustrate our grid of criteria with an itinerary case-study.

2. SPATIAL INFORMATION SUMMARIZATION BASIS
The cumulative approach is a well-known method for summarization in various research areas dealing with space. The GIPSY project [1] is an example of a spatial indexing system (figure 1). Its goal is to find out the most pertinent spatial area within a text document, raising cells of a geo-located grid each time the corresponding spatial information is mentioned in the document. Another example is presented in Information Visualization area (figure 2). The SPIRE project goal [2] is visualization metaphors development. It proposes a 3D landscape where the body of documents is represented by valleys and mountains based on the statistical frequency of key words: the more relevant a document, the higher the mountain that represents it.

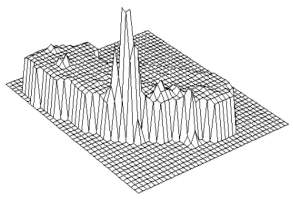


Figure 1. GIPSY project[3]

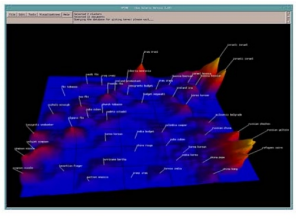


Figure 2. SPIRE project[4]

[1] South-western mountains of France

[2] Syntagm containing spatial information

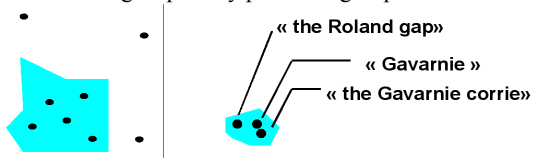
[3] http://www.feweb.vu.nl/gis/SPINlab/education/GIPSY/GIPSYIntroduction.asp

[4] http://www.hipertext.net/english/pag1007.htm

Our approach uses other criteria in order to add semantics to results given by a cumulative approach. We propose a more precise analysis allowing classification of sets of SFs. This analysis is based on SFs and their context interpretation. Therefore spatial information summarization consists in associating "spatial patterns" to text-units.

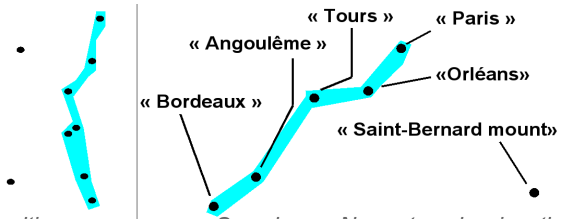
Text-units might have different granularity levels: they correspond to sets of sentences, paragraphs, sections or chapters. We aim at analyzing text-unit SFs in order to point out their prevailing spatial aspect.
Therefore, we may associate a text-unit to one of the three following patterns:

a) A **"point of view"** description (localized description of a town for example), defined by spatial features of various scales forming a spatially pertinent group of areas.



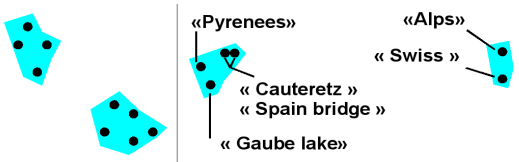
Point of view pattern — *Sample: «Travel in Pyrenees» p 43 - Pimientos.*

b) An **"itinerary"** description (journey narrative with spatial features landmarks), defined by spatial features forming a quite linear or curving geometry and ordered in the text-unit.



Itinerary pattern — *Sample: « New travels in the Pyrenees » p 34 - PyréGraph*

c) An **"area comparison"** between two geographic areas, defined by two spatial features away from each other mentioned in a same text-unit.



Area comparison pattern — *Sample: « New travels in the Pyrenees » p 46 - PyréGraph*

Such aggregations can also be interpreted as complex relative spatial features (R_SF). Therefore, we add three new "point of view", "itinerary", "area comparison" relationships to the core model. In this way, we summarize a text-unit by one (or few) prevailing R_SFs. So, at any level of granularity of the document structure, spatial information may be indexed as instances of our core model.

3. SFs SUMMARIZATION

This section presents the characteristics used to build a spatial interpretation [3,4]. Firstly we use this interpretation to associate a spatial (geo-located) pattern to a set of SFs. This work enables a multi-scale indexing, associating spatial patterns to any kind of text-unit. Then, an itinerary case-study is presented.

3.1 Multi-scale Indexing

Our approach is based on the PIV system spatial information extraction process. PIV indexes contain A_SFs and R_SFs description: name, geo-location, etc. These SFs are extracted from sentence chunks. We want now to abstract/summarize this information in order to build a more advanced multi-scale index using our patterns. Indeed, we make the assumption we can assemble SFs interpretations at a higher level using a function of summarization S: $I_{n+1} = S(I_n)$. I_n represents a text-unit whereas n represents its hierarchical level (sentence, paragraph, section, etc). The S function performs a classification in order to determine for each text-unit the correct pattern, using properties of SFs. These properties are split into two main categories, the ones that refer to SFs themselves (type, scale) and the ones that refer to their disposition (scattering, connection, linearity, distance and salience degree). We use these properties in order to build the main characteristics composing a function of classification. The chart presented in figure 3 lists all the characteristics to be taken into account. They have been drawn from representative samples readings of documents:

Classification	A_{prev}	O_{SFs}	SC_{SFs}	Q_{RSFs}	D_{SFs}	S_{SFs}	Semantic indicators
Itinerary	<40%	>60%	>=50%	>80%	>50%	>60%	
point of view							
area comparison							

Figure 3. Chart presenting characteristics values required for an itinerary hypothesis

a) A_{prev} is the area prevalence: it is the sum of overlapping SFs divided by the sum of SFs. A high percentage increases the importance of a geographic area and allows us to know which one emerges.
In the case of an "itinerary" hypothesis, we don't expect a prevalence for one area. That's why our criterion is less than 40%.

b) O_{SFs} is the SFs "order": this characteristic points out a potentially existing spatial order between SFs. This characteristic is highly relevant for an itinerary hypothesis. The idea is to take the first SF as the starting point, the last SF as the ending point and then to check out if the middle SFs move away from the starting point and approach the ending one.
In the case of an "itinerary" hypothesis, we need more than 60% correctly ordered. We accept a few number of noise.

c) SC_{SFs} is the SFs' scale: it measures the distribution of SFs in scale categories. Indeed, we have split the different kinds of SFs and associated them a scale range. So they can be either microscopic (<25km^2) small (25 to 100 km^2) average (100 to 10000km^2) or big (>10000km^2). The prevalence of one or few of these categories is an indicator to better know which aspect is depicted with a set of SFs.
In the case of an "itinerary" hypothesis, we expect a majority of small and microscopic SFs. So the criterion is at least 50% of SFs in these 2 categories.

d) Q_{RSFs} is the quantity of specific R_SFs: it is the sum of R_SFs with specific relationships divided by the sum of all

R_SFs. Let's note that a list of specific relationships has been associated to each of the 3 patterns.
For an itinerary hypothesis, there are relationships forming geometrical figures ("From A to B", "Between A and B", "the A,B,C triangle", where A, B, C are SFs), inclusion relationships ("crossing A") or close adjacency ("close to A"). We expect a large majority of these R_SFs in comparison with other R_SFs (more than 80%).

e) D_{SFs} is the distance characteristic: it computes specific distance properties according to each pattern hypothesis.
For the itinerary hypothesis, small SFs must be not too far from each other. This assertion is translated by a majority of small SFs (>50%) below the average distance.

f) S_{SFs} is the salience: this last characteristic is highly relevant for the itinerary hypothesis. It consists in computing the average salience between 3 ordered SFs (of the same scale). If the salience is not too high, it can correspond to a way-point. More than 60% of the SFs considered as way-points means an increase of the itinerary weight.

Here is the expression for the weight computing of an itinerary spatial pattern (W_{iti}). Each characteristic can be worth between 0 and 1: $W_{iti} = 1/6(A_{prev}+O_{SFs}+SC_{SFs}+Q_{RSFs}+D_{SFs}+S_{SFs})$. Similar computations are made for the two other patterns, and the heaviest weight helps to choose the better suited pattern.

Indeed, the main idea of our work is to use the spatial scattering of the SFs prior to semantic analysis. However this additional characteristic, based on grammatical relations and verbs analysis, can be useful to validate the hypothesis.

3.2 An Itinerary Case-study

This section presents a text-unit describing an itinerary and explains how pattern characteristics work (figure 3). For now, a set of SFs resulting into an itinerary summary is presented as an example.

3.2.1 Classification:

The following list (figure 4) shows the different spatial features extracted in the narration order.

Figure 4. Representation of the SFs (without Europe(2) and the Alps(5)) according to their order in the text

First of all, in order to determine what kind of spatial pattern we deal with, we have to compute the different patterns weights. We present the itinerary weight computation (W_{iti}) sample only:

a) A_{prev}: the figure 5 shows the maximum accumulation of geographic areas. It represents 5/16 that is to say 31%: "Bordeaux" (twice), "France" and "Europe" overlap on one same area, corresponding to Bordeaux area. Moreover, there is one more SF: "les coteaux de la Bastide" that is a Bordeaux district and corresponds finally to the most often mentioned geographic area.

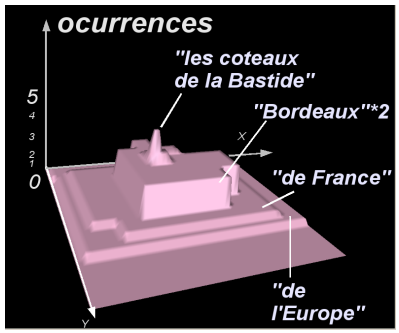


Figure 5. SFs accumulation schema

b) O_{SFs}: we take respectively the first and the last term as the starting and the destination point. If we take into account only the smallest SFs of the text-unit, we have the sub-list: ***Paris,*** *du Mont Saint Bernard, Orléans, Tours, Angoulême, sur les rives de la Dordogne, entre la Garonne et la Dordogne, les coteaux de la Bastide,* ***pont de Bordeaux***. There are 8/9 correct SFs, that is to say 89%.

c) SC_{SFs}: the percentage of the microscopic and the small SFs (cities, hills, etc) is 9/16, that is to say 56%.

d) Q_{RSFs}: "13- on the Dordogne banks" and "14-between the Garonne and the.Dordogne" R_SFs have respectively a geometrical figure relationship and a close adjacency relationship. These relationships are in the "itinerary connoted" list, so we have 100%.

e) D_{SFs}: the histogram hereinafter (figure 6) shows that the majority of our SFs are in the 0-"average distance" interval: 8/14 i.e. 57%.

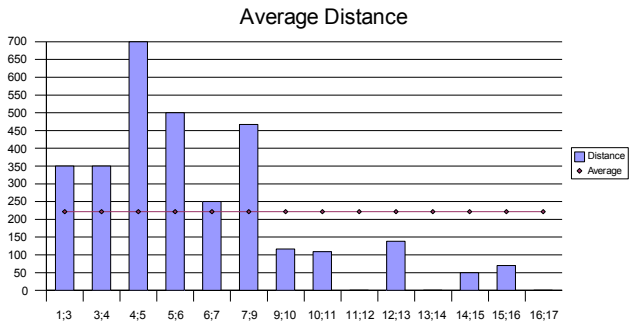


Figure 6. Histogram of the distances between couples of SFs (Orléans-Tours, etc.)

f) S_{SFs}: the computation reveals a salience lower than 60° for the small SFs (Orléans-Tours-Angoulême for instance) except for the "Saint-Bernard mount" SF; therefore we have 8/9 that is to say 89%.

In conclusion, this example has a strong "itinerary weight". An additional semantic analysis may increase this weight: indeed, in this example there are lots of verbs of movement.

Now we have to perform the summarization representation.

3.2.2 Itinerary Computation

In order to give a correct geo-located representation to a text-unit, the S function must use the results coming from the characteristics computation. For an itinerary, the main idea is:

- to take the list of the smallest SFs given during the SC_{SFs} computation ("microscopic" and "small" in our example),
- to filter the too far SFs thanks to D_{SFs} and the ones with a too big salience thanks to S_{SFs} (Saint Bernard Mount),
- then to link the remaining SFs' sub-list geo-located representations (ordered thanks to the list computed at the O_{SFs} step).

GIS functions can "polygonize" them in order to have an itinerary representation (figure 7). Further on we should take into account more SFs, like linear ones (rivers, roads) to compute more complete and precise itineraries.

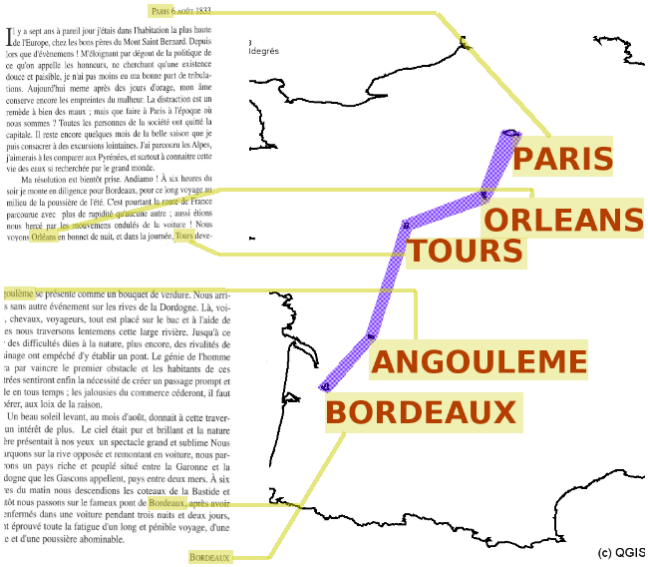


Figure 7. The "Paris-Bordeaux" polygon (ont the right) represents the text-unit (on the left) spatial summary

4. CONCLUSION

In this paper we proposed a pattern based approach for summarizing spatial information. We defined six spatial characteristics and proposed tools to weigh each of them and finally summarize a set of SFs with the prevailing pattern geo-located footprint. This is a prospective work we plan to experiment and extend in the next months.

We use these characteristics for a first level of summarization: text-units of the first hierarchical level in the document structure (paragraph or section). We have to further define summaries of summaries for the higher hierarchical levels. Instead of working with larger sets of SFs, we plan to study spatial patterns summarizing possibilities.

Such summarized spatial indexes should provide new possibilities for spatial IR: faster and smarter spatial criterion based access to paragraphs, sections, chapters of documents, or spatial pattern based querying.

5. ACKNOWLEDGMENTS

Our project is led in partnership with the Pau metropolitan council and the MIDR media library.

6. BIBLIOGRAPHY

[1] Allison Woodruff and Christian Plaunt, GIPSY: Automated Geographic Indexing of Text Documents, Journal of the American Society of Information Science, 1994.

[2] Beth Hetzler, Paul Whitney, Lou Martucci, Jim Thomas, Multi-faceted Insight Through Interoperable Visual Information Analysis Paradigms. In Proceedings of IEEE Symposium on Information Visualization, InfoVis '98, October 19-20, 1998, Research Triangle Park, North Carolina, pp.137-144.

[3] Herzog,G and Maaß,W and Wazinski,P, VITRA GUIDE: Utilisation du Langage Naturel et de Représentation Graphiques pour la Description d'Itinéraires, Images et Langages: Multimodalité et Modélisation Cognitive, Colloque Interdisciplinaire du Comité National de la Recherche Scientifique, Paris, 1993.

[4] A. Klippel and J. Davies and S. Winter and S. Hansen, A High-Level Cognitive Framework for Route Directions, Proceedings of the SSC 2005 Spatial Intelligence, Innovation and Praxis: The National Bienneal Conference of the Spatial Science Institute.

[5] J. Lesbegueries, M. Gaio, P. Loustau, and C. Sallaberry, Geographical information access for non-structured data, ACM SAC ASIIS 2006.

[6] Vandeloise, C. L'espace en français. Travaux de Linguistiques. Seuil, 1986.

[7] M. Hassel 2003. Exploitation of Named Entities in Automatic Text Summarization for Swedish. In the Proceedings of NODALIDA 03 - 14th Nordic Conference on Computational Linguistics, May 30-31 2003, Reykjavik, Iceland

[8] D.A. Randell, Z. Cui and A. Cohn, A Spatial Logic based on Regions and Connection, Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning, pp 165-176, 1992