NORMALIZING SPATIAL INFORMATION TO IMPROVE GEOGRAPHICAL INFORMATION INDEXING AND RETRIEVAL IN DIGITAL LIBRARIES

Damien Palacio and Christian Sallaberry and Mauro Gaio

LIUPPA, Université de Pau
avenue de l'Université
64000 PAU, FRANCE
damien.palacio@univ-pau.fr, christian.sallaberry@univ-pau.fr, mauro.gaio@univ-pau.fr

KEY WORDS: Geographic Information Retrieval, Spatial Information, Normalization, Geographic Information Combination

ABSTRACT:

Our contribution is dedicated to geographic information contained in unstructured textual documents. The main focus of this article is to propose a general indexing strategy that is dedicated to spatial information, but which could be applied to temporal and thematic information as well. More specifically, we have developed a process flow that indexes the spatial information contained in textual documents. This process flow interprets spatial information and computes corresponding accurate footprints. Our goal is to normalize such heterogeneous grained and scaled spatial information (points, polylines, polygons). This normalization is carried out at the index level by grouping spatial information together within spatial areas and by using statistics to compute frequencies for such areas and weights for the retrieved documents.

1 INTRODUCTION

The digitization of printed literature is currently making significant progress. The Google Books Library Project, for instance, aims at creating digital representations of the entire printed inventory of libraries. Other initiatives specialize in the legacy literature of specific domains, such as medicine or cultural heritage (Sautter et al., 2007). For instance, libraries or museums are now offering their electronic contents to a growing number of users.

While some projects only aim at creating digital versions of the text documents, domain-specific efforts often have more ambitious goals (Sautter et al., 2007). For example, to maximize the use of the contents, text documents are annotated and indexed according to domain-specific models. The Virtual Itineraries in the Pyrenees[1] (PIV) project[2] consists in managing a repository of the electronic versions of books (histories, travelogues) from the 19th and 20th centuries. It appears that the contents present many geographic aspects (Marquesuzaà et al., 2005). This kind of repository is quite stable (few suppressions and modifications, regular inserts of documents) and not too large. Therefore, the cost of a back-office refined semantic aware automated indexing is reasonable (Gaio et al., 2008).

Although well-known search engines still deliver good results for pure keyword searches, it has been observed that precision is decreasing, which in turn means that a user has to spend more time in exploring retrieved documents in order to find those that satisfy his information needs (Kanhabua and Nørvåg, 2008). One way of improving precision is to include a geographical dimension into the search. We consider the generally accepted hypothesis that Geographical Information (GI) is made up of three components namely spatial, temporal and thematic. A typical textual sample is: "Fortified towns in the south of the Aquitaine basin in the 13th century." To process this textual unit, we claim that each of its three components (spatial, temporal and thematic) should be treated independently, as is put forth by (Clough et al., 2006). This can be done by making several indexes, one per component, as is advised by (Martins et al., 2005). In this way, one can limit the search to one criterion and easily manage the indexes (e.g., to allow adding documents to the corpus). So, our approach consists in processing components independently, in order to better combine them later on. It contributes to the field of Geographic Information Retrieval (GIR) as defined by (Jones and Purves, 2006).

The current version of the PIV platform is comprised of three independent process flows: spatial (Gaio et al., 2008), temporal (Le Parc-Lacayrelle et al., 2007) and thematic (Sallaberry et al., 2007). For example, Figure 1 illustrates automatic annotations resulting from such process flows: spatial information is highlighted, temporal information is outlined and the thematic one is underlined. Figure 2 illustrates the richness and accuracy of the resulting specific indexes: i.e., the PIV computes geometric representations of spatial information, time intervals corresponding to temporal information and lists of terms corresponding to thematic information. Experiments (Sallaberry et al., 2007) demonstrate the effectiveness of these indexes within specific spatial, temporal or thematic information retrieval scenarios. Two important problems were pointed out during these experiments: 1-results scoring does not integrate spatial features or temporal features frequency within documents: e.g. we are looking for "Biarritz," D1 and D2 will have the same weight even if D1 contains only "Biarritz" spatial feature whereas D2 contains "Biarritz" spatial feature and many other ones; 2-merging results within a geographic information retrieval process remains a challenge (Visser, 2004): as each index is built with one dedicated approach, as well as each document relevancy calculation formula is based on different methods (which correspond respectively to spatial, temporal or thematic criteria), how to combine spatial, temporal and thematic specific relevancy scores of the retrieved documents?

We propose to normalize each geographic indexing criteria. It consists on rearranging geographic information within a uniform representation form: we represent geographic information within spatial tiles (spatial areas), temporal tiles (calendar intervals) and thematic tiles (concepts) and compute each tile evocation frequency in the documents. Then, we apply statistic formulae generally used for plain-text information retrieval to compute relevancy scores for each resulting document.

[1] Mountains of the south west of France

[2] Part of this project is supported by the Greater Pau City Council and the MIDR media library

D1: [...] I visited Biarritz during summer 2000.[...]
D2: [...] Wednesday 16th October 2009
[...] The tramway was often out of order during this week in Bordeaux. [...] I plan to leave Bordeaux next week-end and to go to Biarritz. [...] Saturday, a walk near Bayonne, Sunday, a hike at La Rhune peak as well as at Sare.

Figure 1: Example of automatically annotated textual documents

SF_Id	Doc_Id	Text	Geometry
#1	D1	'Biarritz'	(.122...
...			
#5	D2	'near Bayonne'	(.121...
#6	D2	'La Rhune peak'	(.123...

T_Id	Term	Frequency
#1	'visit'	D1,1;
#2	'tramway'	D2,1; D3,1
...		
#7	'walk'	D2,1; D4,2

TF_Id	Doc_Id	Text	Time interval
#1	D1	'during summer 2000'	06/21/2000-09/21/2000
...			

Figure 2: Example of spatial, temporal and thematic indexes

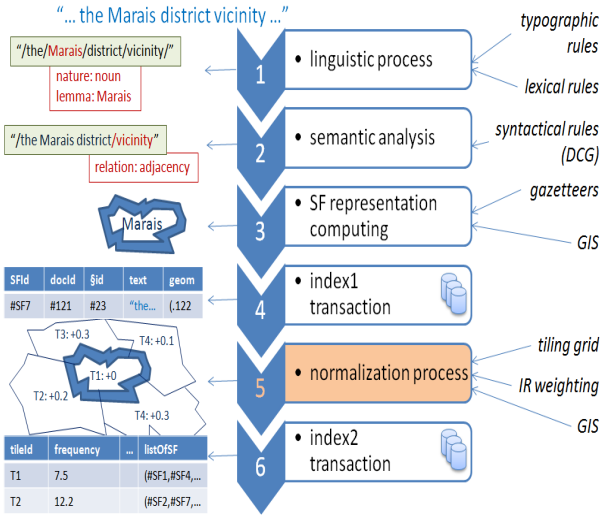

Figure 3: Spatial information process flow

This approach proposes (1) frequency parameter integration within the relevance scoring algorithms and (2) geographic data normalization within a new level of spatial, temporal and thematic indexes. Moreover, we propose to produce different granularity level indexes (for example, spatial administrative segmentations: cities, counties, countries) in order to parse the indexes best suited to the grain of each query.

Merging results provided by such hybrid querying criteria would only make sense if such normalized indexes were homogeneous as well as if the relevance calculation formulae were similar. That is why the next section presents a spatial normalization approach, which we will later apply to the temporal and thematic aspects.

The paper is organized as follows. Section 2 briefly outlines the textual process flow indexing geographic information within the PIV prototype. Section 3 describes related works and our proposals for the creation of new indexes through spatial normalization. Section 4 details the proposed model for computing spatial relevance and describes experiments we carried out to evaluate these propositions. Finally, section 5 and 6 discuss our future perspectives and conclude.

2 TEXTUAL PROCESS FLOW LEADING TO SPATIAL NORMALIZATION

A document textual content processing sequence is usually composed of four main steps: (a) "tokenization" splits the document into smaller blocks of text, (b) lexical and morphological analysis carries out recognition and transformation of these blocks into lexemes, (c) the syntactic analysis, based on grammar rules, allows links between lexeme to be found, finally, (d) the "semantic" step carries out a more specific analysis allowing meaningful lexeme groupings to be interpreted.

As explained hereinafter, our data processing sequence is quite different. This spatial information process flow is described in Figure 3. Steps 1 to 4 are detailed in (Gaio et al., 2008). This approach was developed and experimented within the PIV project:

1. After a classical textual tokenization preprocessing sequence and according to (Baccino and Pynte, 1994) we adopt an active reading behavior, that is to say sought-after information is a priori known. A marker of candidate spatial token locates spatial named entities using typographic and lexical rules (involving spatial features initiator lexicons). Then, a morphosyntactic analyzer associates a lemma and a nature with each candidate token (e.g. "Marais", noun).
2. A semantic analyzer marks candidate Absolute Spatial Features (ASF, e.g. "Marais district") first and candidate Relative Spatial Features (RSF, e.g. "Marais district vicinity") next thanks to a Definite Clause Grammar (DCG). For instance, syntagms of composed nouns (i.e. "Marais district," "Emile Zola street," "Wild Chamois peak") are brought together and spatial relationships (adjacency, inclusion, distance, cardinal direction) are tagged (Egenhofer, 1991).
3. ASF are validated and geolocalized thanks to external and/or internal gazetteers (IGN French Geographic Institute resources, Geonames resources and contributive hand-craft local resources). Then expressions containing RSF are built from pointed out ASF: embedded spatial relationships (e.g. adjacency: "vicinity") are interpreted and corresponding geometries are computed.
4. Only validated spatial features are retained. Thus we get a spatial index describing each SF with the corresponding geometry, text, paragraph and document. This first level of index supports IR scenarios: query/index overlapping geometries are computed and scored relevant textual paragragraphs are returned.
5. SF are grouped, weighted and mapped into a set of segmented grids. We propose different grained tiling grids: regular and administrative grids (district, city, county, ...). We use information retrieval TF.IDF formulae (Spärck Jones, 1972) to compute spatial tiles' frequencies and weight them.
6. Finally, we get a spatial index describing each tile with the corresponding frequency and SF (geometry, text, paragraph and document). This second level of index supports new IR capabilities: a query is mapped to the more convenient grid and query/index overlapping areas are computed and relevance scoring algorithms integrate each tile frequency. This promotes textual paragraphs centered on the required SF only. Moreover, it allows different querying strategies: for example, thin-scale queries are compared to district grids and large-scale ones are compared to country grids.

A GIS supports spatial operations of all the previous stages. This paper focuses on the spatial information normalization process (stage 5). It describes statistical IR approaches integration in such a process. An experiment compares different index tiling grids and IR statistical formulae to validate our propositions.

3 SPATIAL INFORMATION GATHERING FOR SPATIAL NORMALIZATION

3.1 Related works

One of the most popular models developed in textual-based information retrieval research is the vector space model (Salton and McGill, 1983). Using a vector space model, the content of each document can be approximately described by a vector of (content-bearing) terms (Cai, 2002). An information retrieval system stores a representation of a document collection using a document-by-term matrix (Table 1), where the element at (i, j) position corresponds to the frequency of occurrence of term i in the jth document (Manning et al., 2008, Cai, 2002).

$$\begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_n \end{array} \begin{array}{c} \begin{array}{cccc} T_1 & T_2 & \dots & T_t \end{array} \\ \begin{pmatrix} w_{11} & w_{21} & \dots & w_{t1} \\ w_{21} & w_{22} & \dots & w_{t2} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & \dots & w_{tn} \end{pmatrix} \end{array}$$

Table 1: Document-by-T matrix within the vector space model

The vector space model can support selecting and ranking of documents by computing a similarity measure between a document and a query or another document (Salton and McGill, 1983). There are obvious advantages and disadvantages of using vector space model in retrieving geographical information. Vector space model is well accepted as an effective approach in modeling thematic subspace and it allows spatial information to be handled the same way as thematic information (Cai, 2002). (Cai, 2002) proposed to manage place names within a vector space model. Place names are integrated as independent dimensions in a vector space model, whereas in fact, they are points (or regions) in a two-dimensional geographical space. In order to improve such a keyword-based search method, (Cai, 2002) proposed to integrate proper ontologies of places as promoted by (Jones et al., 2001).

Our approach is different as it extends such a term-based matrix to a tile-based matrix. In the vector space model, all the objects (terms, spatial tiles, temporal tiles, thematic tiles (concepts)) can be similarly represented as vectors. This paper proposes to gather SF into spatial tiling grids to compute a similar document-by-tile matrix (Figure 1), where the element at (i, j) position corresponds to the frequency of occurrence of spatial tile i in the jth document.

On the one hand, current spatial oriented research works distinguish:

- spatial generalization: defined as spatial features selection, displacement and/or simplification processes (Zhang, 2005, Zhou and Jones, 2004, Zhou et al., 2000, Glander and Döllner, 2007);
- spatial normalization: defined as an image registration process estimating and applying warp-fields (Robbins et al., 2003);
- spatial summarization: defined as spatial features aggregation / combination into larger features (i.e. cell-based structure) (Rees, 2003).

On the other hand, information retrieval oriented research works define normalization as a stemming process of words in order to gather and weight them (Spärck Jones, 1972, Li et al., 2002). So, what we call normalizing spatial information, in the following section, means spatial information (representations computed from textual documents) gathering into spatial tiles in order to weight them according to frequency computations.

The originality of the approach described in the following section consists in:

a) the proposition of different granularity level spatial indexes: administrative and/or regular grids;

b) the adaptation of effective full-text IR technics in order to process such indexes.

3.2 Spatial Gathering for normalization

First, we detail the spatial normalization process (stage 5 Figure 3) leading to the index2 (stage 6 Figure 3). Then, we briefly explain how we take advantage of this normalized index within an IR process.

3.2.1 Information Indexing. Our approach consists in gathering spatial information into a unique type of spatial representation: the tile. So we divide space by attaching each detected SF to tiles. It is similar to the lemmatisation process, for which each term is attached to a lemma. Two segmentation types are possible. The first concerns regular tiles (i.e., segmentation into rectangular tiles of the same size — see Figure 4). It is similar to truncation [3]. The second concerns administrative tiles (i.e., segmentation into cities for example — see Figure 5). It is similar to lemmatisation [4]. To calculate a tile frequency, one just has to count the number of SF that intersect it, while keeping in mind that a SF can intersect several tiles.

For illustration purposes, let's go back to the example in Figure 1 and 2. If we choose to use regular segmentation (Figure 4), we obtain the tiles index shown in Table 2. In this table, several scenarios are presented. First, SF #5 intersects two tiles (T2 and T3); so the discrete frequency of both of them is incremented by 1. Moreover tile T2 is intersected by two SF (#1 and #5); consequently it has a weight of 2.

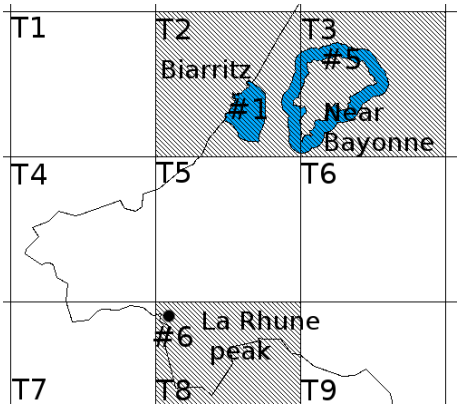

Figure 4: Part of thin-grained SF obtained in index1 projected on a segmentation by regular grid

id_t	id_{sf} list	discrete frequency	continuous frequency
T1	[]	0	0
T2	[#1;#5]	2	0.15
T3	[#5]	1	0.20
...			

Table 2: Spatial index2 with regular tiles (phase 6 - Figure 3)

One should note the granularity problem of the spatial information that is being processed, and the proportionality issue between

[3] e.g. for word "forgotten" the truncation returns "forgott"

[4] e.g. for word "forgotten" the lemmatisation returns "forget"

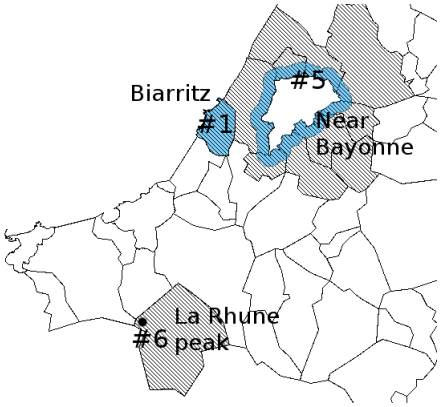

Figure 5: Part of thin-grained SF obtained in index1 projected on an administrative segmentation (cities)

Discrete frequency	$freq_t = \sum freq_{sf}$
Continuous frequency	$freq_t = \sum freq_{sf} * \frac{Ar_{sf,t}}{Ar_t} * \frac{1}{NbTiles_{sf}}$

Table 3: Frequency formulae ($Ar_{sf,t}$: SF area on tile t, Ar_t: tile t area, $NbTiles_{sf}$: number of tiles intersected by the SF)

its representation (SF) and a tile as well as the size of their overlapping area. Indeed, one may wonder whether a SF's area that only covers a small part of a tile should have as great a weight as a SF's area that covers most of the area of the same tile? Thus, we suggest two frequency calculation (see Table 3). Indexing may be discrete; so, for a given document unit, a tile frequency is incremented by 1 every time there is an intersection with a SF (Table 2 column 3). We have also considered a continuous indexing approach. According to the ratio overlay SF/tile, a tile frequency is incremented by a value between 0 and 1 (Table 2 column 4).

These indexes are intended to support spatial IR. This involves weighting the results. Consequently, we use regular IR formulae and carry out experiments on such indexes of spatial tiles.

3.2.2 Information Retrieval. We propose 4 IR formulae (Table 4). TF, TF.IDF and OkapiBM25 are dedicated to discrete weighting. They are widely used in full-text IR (Manning et al., 2008, Savoy, 2002). In our context, the TF formula avoids reducing the weight of overly frequent tiles. Nevertheless, classical frequency does not take spatial specificity like granularity into account. That is why we decided to apply TF onto continuous frequency, and we call this approach TFc.

4 EXPERIMENTS

Our hypothesis is that the segmentation must be adapted to the SF type of the corpus and of the queries. We propose to use the granularity of the query to choose the best suited index. For complex queries, composed of different grained SF, we propose to define a default well suited index. So here we are looking for what segmentation and weighting formula give the better results for our corpus.

Our approaches are based on thin-grained spatial data. But spatial evaluation campaigns like GeoCLEF (Mandl et al., 2007) do not give accurate resources (like polygons) and do not handle French documents. That's why we realize our experiment on our cultural heritage digital library.

10 French books were indexed. In order to compare our propositions to manually sorted methods, we chose a sample of 1,019

Tile Frequency (TF)	$W_{t,Du} = TF_{t,Du} = \frac{freq_{t,Du}}{\sum_{i=1}^{n} freq_i}$
TF.IDF	$W_{t,Du} = TF_{t,Du} * IDF_t$ and $\text{IDF}_t = log\left(\frac{NDu}{NDu_t}\right)$
OkapiBM25	$W_{t,Du} = \left(\frac{(k_1+1)*TF_{t,Du}}{(K+TF_{t,Du})}\right)$ and $\text{K} = \text{k}_1 * [(1-b) + \frac{b*n}{advl}]$
TFc	$W_{t,Du} = TFc_{t,Du} = \frac{freqC_{t,Du}}{\sum_{i=1}^{n} freqC_i}$
$freq_{t,Du}$: tile frequency in the document unit, n : number of tiles in the document unit, NDu_t : number of document units with tile t, NDu : number of document units, k_1 = 1.2, b = 0.75, $advl$ = 900, $freqC$: continuous frequency	

Table 4: Weighting formulae, used with index2, for a tile t and a document unit Du

in downtown Paris (Inclusion)	near Gavarnie (Proximity)
on Tarbes-Lourdes axis (Union)	in south of Ile-de-France (Orientation)

Table 5: Examples of RSF

document units, corresponding to 1,028 SF (902 ASF and 126 RSF). Each document unit may contain from 0 to many SF. We submitted 40 queries (the baseline is index1). 15 queries involve an ASF : 5 of each type (small grained like peaks, intermediate grained like cities and large grained like regions). 25 other contain a RSF : 5 of each type (orientation, proximity, union, inclusion, distance). Table 5 shows some examples of relations. We observed that 30% of our ASF are well identified cities, 12% are larger well identified ASF (department, regions), 38% are smaller ASF (peaks, cabins, ...) and the others (approximatively 20%) have a variable average size.

We tried 6 different indexing segmentations: 3 administrative segmentations (city, department and region) and 3 regular segmentations (grid of 100x100, grid of 200x200 and grid of 400x400). The grid of 200x200 corresponds to the average city size. Finally, we tested all theses segmentations with the 4 weighting formulae presented in last section (TF, TF-IDF, OkapiBM25, TFc).

As we can see in Table 6, for all segmentations, the TFc gives the best results. As we explained in section 3, every classical statistical weighting formulae (TF, TF-IDF, OkapiBM25) use discrete frequency. They give the same weight for a geometry which fills the major part of one tile, and for a geometry which represents a little part of the same tile. On the contrary, the TFc uses continuous frequency and gives a weight depending on the area of overlapping between the tile and the SF's geometry.

Concerning the segmentation, the Table 6 shows that all segmentations give good results excepts department and regions (they are too large so they gather SF which are too far away from each other). For segmentation by regular grid, the one of 200x200 gives the best results. Concerning the administrative segmentation, city segmentation gives the best results. The main explanation is that an important part of the indexed ASF concerns well identified cities. So it confirms our hypothesis that the segmentation must be adapted to the type of the SF contained in the corpus.

Tables 6 and 7 also show that the city segmentation associated to the TFc gives better results than our baseline. Let's take example of Figure 1 to illustrate why we obtain such results. If we consider query "in Biarritz," the relevancy score for D2 on index1 is 1.0 because the text contains the SF "in Biarritz." It does not take into account the other SF. On the other hand, the city segmentation associated to the TFc gives a relevancy score of about 0.17. It computes a lower score to the document unit because it contains other less relevant SF.

MAP	TF	TF-IDF	Okapi	TFc
City Segmentation	0.61	0.61	0.63	**0.70**
Department Segmentation	0.40	0.39	0.40	**0.53**
Region Segmentation	0.40	0.39	0.39	**0.56**
Grid of 100x100	0.59	0.59	0.62	**0.68**
Grid of 200x200	0.61	0.60	0.63	**0.69**
Grid of 400x400	0.63	0.62	0.65	**0.66**

Table 6: Results of experiment on SF with different segmentations and weighting formulae

MAP	Spatial Overlapping
index1 (baseline)	0.62

Table 7: Results of experiment on SF with baseline (index1)

In conclusion, we advise segmentation into cities and the TFc formula (cf Table 6) for cultural heritage digital libraries. This normalization allows one to introduce an initial approximation of the spatial context (weighting a document unit takes into account all the SF it contains).

5 ONGOING AND FUTURE WORK

The PIV platform supports a similar processing sequence producing temporal indexes (Figure 3). It deals with calendar temporal features (CTF) that may be absolute (ACTF) or relative (RCTF) like spatial ones.

Let one consider that the previous text sample involves the following temporal features: CTF1-"the 26th of December", CTF2-"Saturday 29th of December at 2pm", CTF3-"at the beginning of the winter", CTF4"the last days of December 1933". The PIV produces such an index (Table 8, Figure 6).

The PIV temporal information normalization process (ongoing development similar to the spatial normalization process) would return weighted temporal intervals presented in Figure 7. This example illustrates calendar segmentation where each interval represents a week: TF4 intersects weeks W51 and W52. For example the week W52 has a weight of 4 according to the discrete indexing approach.

Currently, we are working on temporal normalization experiment. We aim to propose spatial and temporal criteria combination strategies with geographic IR scenarii.

id_{ctf}	text	type	timestamp
ctf1	the 26th of December	actf	(1933-12-26, 1933-12-26)
ctf2	Saturday 29th of December	actf	(1933-12-29, 1933-12-29)
ctf3	at the beginning of the winter	rctf	(1933-12-22, 1934-03-19)
ctf4	the last days of December 1933	rctf	(1933-12-21, 1933-12-31)

Table 8: Index of temporal features in PIV

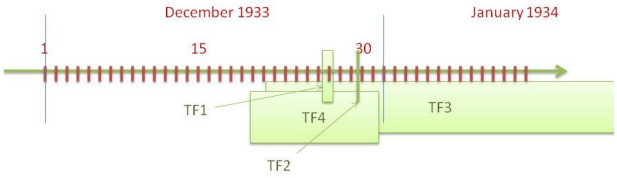

Figure 6: Temporal Index

6 CONCLUSION

The Virtual Itineraries in the Pyrenees (PIV) project consists in managing a repository of the electronic versions of books (histories, travelogues) from the 19th and 20th centuries. The PIV engine automatically annotates, interprets and indexes spatial, temporal and thematic information contained in those documents. Three independent process flows support spatial, temporal and thematic indexing and IR operations.

Two important problems were pointed out during a first campaign of experiments (Sallaberry et al., 2007): 1-results scoring does not integrate spatial or temporal features frequency within documents; 2-merging results within a geographic information retrieval process remains a challenge. The main problem of current geographic IR systems comes from the fact that the index structure and relevancy computation approaches used for space, time and theme are intrinsically different (Visser, 2004).

Our hypothesis is based on a spatial, temporal and thematic tiling of information in order to build higher level indexes and to adapt effective full-text IR technics to process such indexes.

In this paper we propose an approach for normalizing spatial indexes automatically. Such a gathering of spatial features into spatial tiles implies some loss of accuracy. However, as we have different grained indexes, we may select the best suited one during a querying stage. Moreover, experiments point out the effectiveness of our solution: a continuous spatial tiles frequency computation associated to a continuous document units relevancy computation formula gives better results than our baseline dedicated to the weighting of the most relevant SF of a document unit.

As explained before, the aim of this normalization method is to develop a general indexing strategy that is suited for spatial, temporal and thematic information in order to combine such geographic IR results. We are currently working on the evaluation of the effectiveness of this indexation on a larger sample of texts and queries. We are also working to apply normalization methods for the building of normalized temporal and thematic indexes from textual input. Future improvement of the presented approach would be to explore how to combine normalized spatial, temporal and thematic indexes and compute a unique relevancy scoring. Merging results for a geographic IR approach combining such different criteria is a recurring research question nowadays (Martins et al., 2008, Vaid et al., 2005).

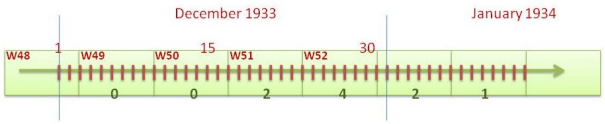

Figure 7: Calendar Segmentation

REFERENCES

Baccino, T. and Pynte, J., 1994. Spatial coding and discourse models during text reading. Language and Cognitive Processes 9, pp. 143–155.

Cai, G., 2002. GeoVSM: An Integrated Retrieval Model for Geographic Information. In: Max J. Egenhofer and David M. Mark (ed.), GIScience, Lecture Notes in Computer Science, Vol. 2478, Springer, pp. 65–79.

Clough, P., Joho, H. and Purves, R., 2006. Judging the Spatial Relevance of Documents for GIR. In: ECIR'06: Proceedings of the 28th European Conference on IR Research, Lecture Notes in Computer Science, Vol. 3936, Springer, pp. 548–552.

Egenhofer, M. J., 1991. Reasoning about Binary Topological Relations. In: Oliver Günther and Hans-Jörg Schek (ed.), SSD, Lecture Notes in Computer Science, Vol. 525, Springer, pp. 143–160.

Gaio, M., Sallaberry, C., Etcheverry, P., Marquesuzaa, C. and Lesbegueries, J., 2008. A global process to access documents' contents from a geographical point of view. In: Journal of Visual Languages And Computing, Vol. 19number 1, Academic Press, Inc., Orlando, FL, USA, pp. 3–23.

Glander, T. and Döllner, J., 2007. Cell-based generalization of 3D building groups with outlier management. In: Hanan Samet and Cyrus Shahabi and Markus Schneider (ed.), GIS, ACM, p. 54.

Jones, C. B., Alani, H. and Tudhope, D., 2001. Geographical Information Retrieval with Ontologies of Place. In: D.R. Montello (ed.), Conference on Spatial Information Theory - (COSIT 2001), Vol. 2205 / 2001, Springer-Verlag Heidelberg, Morro Bayand California USA, pp. 322–335.

Jones, C. B. and Purves, R., 2006. GIR'05 2005 ACM workshop on geographical information retrieval. SIGIR Forum 40(1), pp. 34–37.

Kanhabua, N. and Nørvåg, K., 2008. Improving Temporal Language Models for Determining Time of Non-timestamped Documents. In: ECDL'08: Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries, Springer-Verlag, Berlin, Heidelberg, pp. 358–370.

Le Parc-Lacayrelle, A., Gaio, M. and Sallaberry, C., 2007. La composante temps dans l'information géographique textuelle. Revue Document Numérique 10(2), pp. 129–148.

Li, H., Srihari, K. R., Niu, C. and Li, W., 2002. Location Normalization for Information Extraction. In: 19th International Conference on Computational Linguistics (COLING2002)- Howard International House and Academia Sinicaand Taipeiand Taiwan, Association for Computational Linguistics.

Mandl, T., Gey, F. C., Nunzio, G. M. D., Ferro, N., Larson, R., Sanderson, M., Santos, D., Womser-Hacker, C. and Xie, X., 2007. GeoCLEF 2007: The CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In: Carol Peters and Valentin Jijkoun and Thomas Mandl and Henning Muller and Douglas W. Oard and Anselmo Penas and Vivien Petras and Diana Santos (ed.), CLEF, Lecture Notes in Computer Science, Vol. 5152, Springer, pp. 745–772.

Manning, C. D., Raghavan, P. and Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, New York.

Marquesuzaà, C., Etcheverry, P. and Lesbegueries, J., 2005. Exploiting Geospatial Markers to Explore and Resocialize Localized Documents. In: M. Andrea Rodríguez and Isabel F. Cruz and Max J. Egenhofer and Sergei Levashkin (ed.), GeoS, Lecture Notes in Computer Science, Vol. 3799, Springer, pp. 153–165.

Martins, B., Manguinhas, H. and Borbinha, J. L., 2008. Extracting and Exploring the Geo-Temporal Semantics of Textual Resources. In: ICSC, IEEE Computer Society, pp. 1–9.

Martins, B., Silva, M. J. and Andrade, L., 2005. Indexing and ranking in Geo-IR systems. In: GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval, ACM, New York, NY, USA, pp. 31–34.

Rees, T., 2003. "C-Squares", a New Spatial Indexing System and its Applicability to the Description of Oceanographic Datasets. In: Oceanography, Vol. 16number 1, pp. 11–19.

Robbins, S., Evans, A. C., Collins, D. L. and Whitesides, S., 2003. Tuning and Comparing Spatial Normalization Methods. In: Randy E. Ellis and Terry M. Peters (ed.), MICCAI (2), Lecture Notes in Computer Science, Vol. 2879, Springer, pp. 910–917.

Sallaberry, C., Baziz, M., Lesbegueries, J. and Gaio, M., 2007. Towards an IE and IR System Dealing with Spatial Information in Digital Libraries – Evaluation Case Study. In: ICEIS'07: Proceedings of the 9th International Conference on Enterprise Information Systems, pp. 190–197.

Salton, G. and McGill, M. J., 1983. Introduction to Modern Information Retrieval. McGraw-Hill.

Sautter, G., Böhm, K., Padberg, F. and Tichy, W. F., 2007. Empirical Evaluation of Semi-automated XML Annotation of Text Documents with the GoldenGATE Editor. In: ECDL'07: Proceedings of the 11th European Conference on Digital Libraries, LNCS, Vol. 4675, Springer, pp. 357–367.

Savoy, J., 2002. Morphologie et recherche d'information. Technical report, Institut interfacultaire d'informatique, Université de Neuchâtel.

Spärck Jones, K., 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28(1), pp. 11–21.

Vaid, S., Jones, C. B., Joho, H. and Sanderson, M., 2005. Spatio-textual Indexing for Geographical Search on the Web. In: Claudia Bauzer Medeiros and Max J. Egenhofer and Elisa Bertino (ed.), SSTD, Lecture Notes in Computer Science, Vol. 3633, Springer, pp. 218–235.

Visser, U., 2004. Intelligent Information Integration for the Semantic Web. Springer-Verlag, Heidelberg.

Zhang, Q., 2005. Road Network Generalization Based on Connection Analysis. In: Developments in Spatial Data Handling, Springer Berlin Heidelberg, pp. 343–353.

Zhou, S. and Jones, C. B., 2004. Shape-Aware Line Generalisation With Weighted Effective Area. In: Developments in Spatial Data Handling 11th International Symposium on Spatial Data Handling, Springer, Springer, pp. 369–380.

Zhou, X., Zhang, Y., Lu, S. and Chen, G., 2000. On Spatial Information Retrieval and Database Generalization. In: Kyoto International Conference on Digital Libraries, pp. 380–386.