TOWARDS AN IE AND IR SYSTEM DEALING WITH SPATIAL INFORMATION IN DIGITAL LIBRARIES – EVALUATION CASE STUDY

Christian Sallaberry*, Mustapha Baziz* **, Julien Lesbegueries* and Mauro Gaio*
** Laboratoire d'Informatique-Université de Pau (UPPA, France)*

*** Institut de Recherche en Informatique de Toulouse (IRIT), France*

Keywords: Geographic Information Extraction and Retrieval, Spatial Information Scope, Classical IR, Digital Libraries.

Abstract: This paper deals with spatial Information Extraction (IE) and Retrieval (IR) in Digital Libraries environments. The proposed approach (implemented within PIV[1] prototype) is based on a linguistic and semantic analysis of digital corpora and free text queries. First, we present requirements and a methodology of semantic annotation for automatic indexing and geo-referencing of text documents. Then we report on a case study where the spatial-based IR process is evaluated and compared to classical (statistical-based) IR approaches using first pure spatial queries and then more general ones dealing with both spatial and thematic scopes. The main result in these first experiments shows that combining a spatial approach with a classical (statistical-based) IR one improves in a significant way retrieval accuracy, namely in the case of general queries.

1 INTRODUCTION

Geographically related queries form nearly one fifth of all queries submitted to Excite search engine, the terms occurring most frequently being place names (Sanderson and Kohler, 2004). Our contribution focuses on digital libraries and proposes to extend basic services of existing Library Management System with new ones dedicated to geographic information extraction and retrieval (PIV project (Lesbegueries et al., 2006)). Geographic information in such a repository is composed of a spatial feature, a temporal feature and a thematic one. "Music instruments in the vicinity of Laruns in the XIXth century" is an example of a complete geographic feature: "Music instruments" is the thematic feature, "vicinity of Laruns" is the spatial feature and "XIXth century" is the temporal one.

Let's assume that to initiate a geographical retrieval process the spatial feature has to be explicit whereas the temporal one could be implicit or not locally expressed and the thematic feature can be missing. Consequently, to process geographical information in-depth, analysis of spatial information is mandatory.

Our spatial model supports absolute and Relative Spatial Features. Spatial features such as "Biarritz district" are well-known named places. We call them Absolute Spatial Features (ASF). Complex Spatial Features as "Biarritz vicinity" or "South of Biarritz district" have to be interpreted and, therefore, need some spatial reasoning processes (Cohn and Hazarika., 2001). Such features are called Relative Spatial Features (RSF). We associate each RSF to one or more spatial relationships (adjacency, inclusion, distance, orientation) for a recursive definition.

Works like the SPIRIT project, the Geosearch system, the GEO-IR system, etc. are related to spatial information management. They are presented in (Chen et al., 2006). A difference of our approach with other ones like SPIRIT (Jones et al., 2004) and GIPSY (Woodruff et al., 1994) relies on the back-

[1] PIV: project named Virtual Itineraries in Pyrenees (mountains of the south-west of France)

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 *Proceedings of the Ninth International Conference on Enterprise Information Systems - HCI*, pages 190-197
DOI: 10.5220/0002383701900197

office spatial reasoning used for both ASFs and RSFs interpretation and indexing. For instance, the SPIRIT system mainly tags ASFs. Another specificity concerns the granularity level of the managed information units: textual paragraphs of a domain specific corpora (cultural heritage of Pyrenees) in our case and web pages in the case of SPIRIT system. In the proposed approach, a refined spatial information interpretation and a markup process are applied both within the information units indexing stage and the users' query interpretation. As we work on specific digital library collections and as these collections are quite stable and not too large, the hard back-office spatial process seems to be suitable (Lesbegueries et al., 2006). Therefore, the cost of such refined spatial aware indexing is reasonable. Queries are interpreted dynamically in the same way and SFs blow-by-blow indexes allow a more accurate information retrieval.

The paper is organized as following. In the second section we present PIV spatial semantics processing. In the third section, we experiment and present the first results of an evaluation and combination of PIV spatial approach with classical statistical IR approaches.

2 PIV PROJECT

2.1 An Overview of the System

In PIV project, we want a non-expert user (tourist, scientist or scholar) to access to territorial-oriented digitized corpora. Figure 1 represents PIV system's two main sub-processes of Information Extraction and Retrieval.

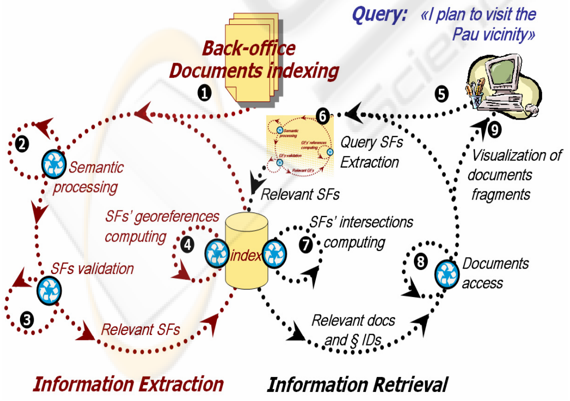


Figure 1: Synoptic schema of information extraction, retrieval and visualization in PIV system.

Roughly, IE is held in four main stages. First of all, documents collections are built (stage (1)), in this paper, we used digitized archives dealing with the cultural heritage of the south west of France. Then in stage (2), a linguistic and semantic analysis of these digital corpora is carried out in order to extract SFs as formal representations of instances of the PIV spatial model. The third stage (3) parses geographic gazetteers (districts, named-places, roads, cliffs, valleys, …) in order to validate SFs captured before. IE then computes spatial representations and georeferences (stage (4)). Thus, the IE sub-processes results are either absolute (e.g. "Laruns village") or relative SFs (e.g. "Laruns village vicinity").

IR part is also based on such an analysis of the query (stage (6)) and relies on a spatial mapping. It computes intersection surfaces (stage (7)) between spatial representations corresponding to the query and those contained in the indexes (cf. §2.4). It will be then necessary to extract fragments of such relevant documents (stage (8)) and, finally, to present them to the user (stage (9)).

2.2 The Spatial Core Model

In this model, according to the linguistic hypothesis, a SF is recursively defined from one or several other SFs and spatial relations are part of the SFs' definition (Lesbegueries et al., 2006, 2006b). The target/landmark principle (Vandeloise 1986) can be defined in a recursive manner. For instance, the SF "north of the Biarritz-Pau line" is first defined by "Biarritz" and "Pau" landmarks that are well known named places, the term "line" creates a new well-known geometrical object linking the two landmarks and cutting the space into two sub-spaces, finally, an orientation relation creates a reference on the target to focus on.

Figure 2 shows that a SF has at least one representation (A) with a natural or artificial boundary; it can be specialized (B) into an absolute (ASF), i.e. "Laruns village" named place or a relative feature (RSF). A RSF is defined with a reference, i.e. "west of Laruns village" relation linking at least one other SF (C). The cycle represents the recursive definition.

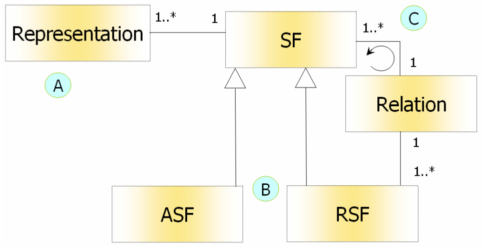


Figure 2: Spatial core model simplified schema.

For spatial information extraction in textual documents, a Definite Clause Grammar illustrated in (Lesbegueries et al., 2006) specifies lexicons and rules in order to detect SFs and create instances of this model.

Thus, a SF spatial relation can be an adjacency ("nearby Laruns"), an inclusion ("centre of Laruns"), a distance ("at about 10 kms of Laruns"), a geometric form ("the Laruns Arudy Mauleon triangle") or an orientation ("in the west of Laruns").

In the core model all of these spatial references have attributes used to characterize them. So, for instance, distance has a numerical and/or a qualitative parameter and adjacency has a qualifier as defined in (Lesbegueries et al., 2006b) and (Muller 2002).

So, a XML tree (cf. §2.3) complying with the PIV XML schema (Lesbegueries et al., 2006) describes any SF.

2.3 Spatial IE and Indexing

Hereinafter, we briefly describe the Linguistic and Semantic Processing Sequence supporting PIV spatial IE process (Lesbegueries et al., 2006).

The LPS goal is to populate a structured information repository (XML indexes) from heterogeneous information sources (news papers and books contents, postcards descriptors). We also used it to separate spatial features from the thematic ones in the query when evaluating IR results (cf. §3.5).

According to works on textual documents (Lesbegueries et al., 2006b), we adopt an active reading behaviour, that is to say sought-after information is known a priori. This is why, unlike slight Natural Language Processing (NLP) (Abolhassani et al., 2003), our linguistic and semantic processing sequence is locally applied near candidates for named places. To mark these candidates a lexicon is used in order to have a quite good generic bootstrap process. So ASFs (i.e. villages' names, forests' names, etc.) are detected first and marked. Then RSFs are built from previously pointed out ASFs. The data processing sequence used to highlight spatial features is implemented as described in Figure 3.

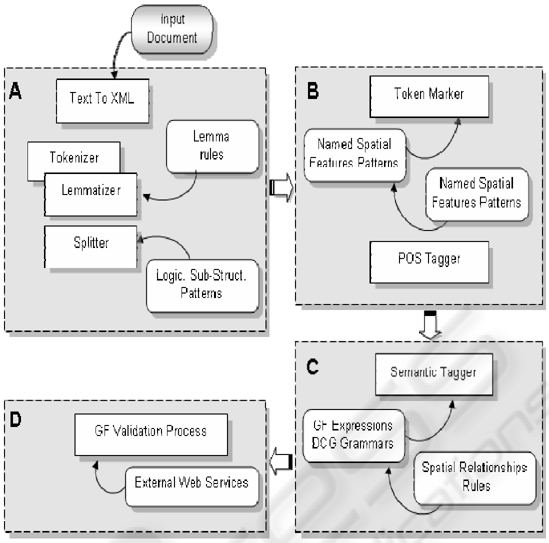


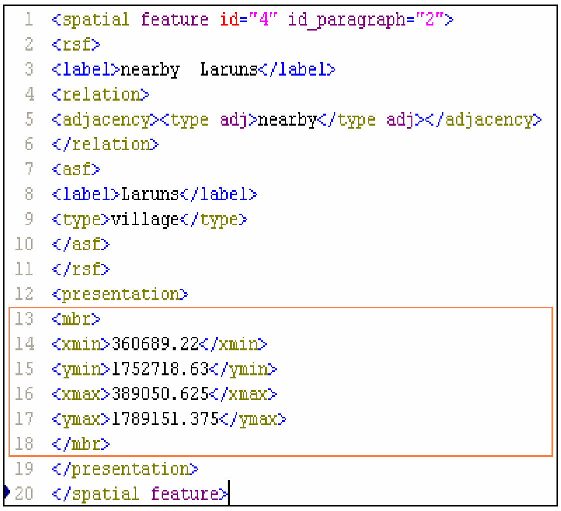
Figure 3: Linguistic/Semantic Processing Sequence (LPS).

First a tokeniser and a splitter parse the textual flow (Figure 3-A). This pre-treatment corresponds to new textual flow where the initial content is added with logical sub-structures marks; words separators marks are added with their lemmas (thanks to a lemmatization phase embedded).

In the second stage (Figure 3-B), spatial features called "candidates" are detected as following: first, all sentences having tokens starting with a capital letter and preceded with a token containing terms specified in a lexicon "in", "from", … (known as spatial feature's initiator) are marked. Then, a Part Of Speech (POS) tagger parses these marked sentences and retrieves words' POS.

In the third stage (Figure 3-C), a Definite Clause Grammar (DCG) based analysis interprets the extracted syntagms (inclusion, adjacency, distance to another spatial feature, etc.). The feature "near of Laruns" is interpreted as a RSF ("rsf" tag in line 2 Figure 4) itself defined by an adjacency relation (line 4-6 Figure 4) and by the "Laruns" ASF (line 7-10 Figure 4).

The SFs validation stage calls external services (gazetteers) to confirm every candidate ASF (Figure 3-D). For the sentence "Paul passe près de Laruns" (Paul passes nearby Laruns): "Laruns" candidate SF is confirmed whereas "Paul" candidate SF is removed. All the RSFs candidates associated to a non-validated ASF are also removed. Finally a MBR (Minimum Bounding Rectangle) (Lesbegueries et al., 2006) representation consisting on geocode coordinates (lines 13-18 Figure 4) is added to the XML index tree.



```
<spatial feature id="4" id_paragraph="2">
<rsf>
<label>nearby  Laruns</label>
<relation>
<adjacency><type adj>nearby</type adj></adjacency>
</relation>
<asf>
<label>Laruns</label>
<type>village</type>
</asf>
</rsf>
<presentation>
<mbr>
<xmin>360689.22</xmin>
<ymin>1752718.63</ymin>
<xmax>389050.625</xmax>
<ymax>1789151.375</ymax>
</mbr>
</presentation>
</spatial feature>
```

Figure 4: An excerpt of the SFs XML indexes.

2.4 Spatial IR Based on SFs Intersections

We use SFs indexes to undertake queries and retrieve information from documents.

A free text interface supports the IR stage. Queries are analyzed exactly as the documents of the corpus are: the same IE data processing sequence is executed and every SF is extracted. All the validated SFs are geo-localized and a MBR is attached to each one of these SFs. A query is analyzed online whereas corpus documents are analyzed offline.

Our search technique is based on a spatial mapping between the SFs of the query and those of the documents (stage (7) in Figure 1). This mapping is done thanks to the geospatial footprints created dynamically for the query and those stored in index files of the corpus.

For example, Figure 5 illustrates a query and an indexed area (precise geospatial footprints for ASFs and approximated MBRs for RSFs).

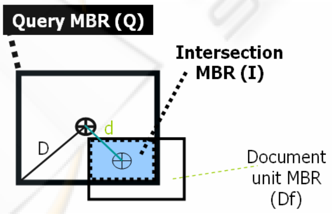


Figure 5: Relevance computing.

The selection process consists in processing index files and computing intersections with a GIS (Lesbegueries et al. 2006). Then, we select corresponding relevant Documents fragments (Df).

We are able to calculate the relevance of a document fragment by computing an evaluation of the surface which results from the intersection between the SF of the document fragment and the ones of the query:

For any query, the relevance of each recovered document may be different (Figure 5):

$$Df\ precision = \frac{I\ surface}{Df\ surface}$$

$$Df\ significance = \frac{I\ surface}{Q\ surface}$$

$$Df\ distance = \frac{d}{D}$$

Therefore, we compute Df score as following:

$$Df\ score = \frac{(Df\ precision + Df\ significance)}{(2 + Df\ distance)} \qquad (1)$$

The closer the centroids of I and Q are to each other, the higher the relevance score of Df.

An XML DBMS (*eXist* - http://exist.sourceforge.net) and a GIS (*PostGIS* - http://postgis.refractions.net) support these searching and computing operations on the corpus indexes. Figure 6 illustrates relevance computing via functions and queries submitted to the GIS.

area(intersection(Q_geom, Df_geom))	I_surface
area(Df_geom)	Df_surface
distance(centroid(Q_geom), centroid(Df_geom))	d
distance(centroid(Q_geom), geomfromtext('corner coordinate'))	D

SELECT pi.gid, pi.doc_name, pi.par_id, pi.SF-name,
(tq.isurf/tq.dfsurf + tq.isurf/tq.qsurf)/(2 + tq.d/tq.D)
AS weight
FROM piv_index pi, temp_query tq
WHERE pi.gid=tq.gid ORDER BY weight DESC;	

Figure 6: Surfaces, distances and score computing.

The query of Figure 6 returns the relevant documents and paragraphs IDs. Then the original texts and the SFs details may be presented in a weighted order.

3 CASE STUDY

In this section, we evaluate the PIV spatial-based IR approach based on information extraction (IE) of Spatial Features (SFs) in textual documents. The PIV results are compared to those obtained by a classical keywords-based IR using the same collection and the same set of test queries. The used classical IR approach is defined in the next section.

3.1 Classical IR Approach

The IR classical approach is based on the notion of "bag" of single words (Baeza-Yates et al., 1999). In such full text approaches, documents are first indexed using a classical term indexing. It consists in selecting single words occurring in the documents, and then stemming these words using an appropriate stemmer (Porter 2001) and at the end removing stop-words according to a stoplist. We used in this paper a stoplist and a French stemmer from the Snowball family of stemmers (Porter 2001). A weight Wtd(t,d) is then assigned to each term t in a document dj following the formula given in (2):

$$Wtd(t_j, d_j) = \frac{\dfrac{2.tf_{ij}.\log(N - n_i + 0.5)}{(n_i + 0.5)}}{2.(0.25 + 0.75\ .\ dl_j / avg\_dl) + tf_{ij})} \quad (2)$$

Where $tf_{i,j}$ represents the frequency of the term t_i in the document d_j, n_i is the number of documents containing the term t_i and N the total number of documents in the collection. dl_j represents the length of the document d_j and *avg_dl,* the average length of the document in the collection. This weighting method, which is an enhanced TF.IDF formula, is introduced to attenuate the negative impact of large documents in the searching stage (Robertson et al., 1995). This is also suitable for the used collection (paragraphs with various lengths). The same indexing process is applied to queries.

A vector-based model (Boughanem et al., 2001) is then used to retrieve documents: for a given query *q*, the Inner product between the vector of the query and the ones of each document d_j in the collection is applied in order to compute the relevance score:

$$\mathrm{Re}\,l(q, d_j) = \sum_{k=1}^{|q|} Wtq(t_k, q).Wtd(t_k, d) \quad (3)$$

Finally, this relevance score is used to determine the ranking of the document (d_j) in the final list of retrieved documents in response to the query (*q*).

3.2 Sample Data

The corpus used for training and testing the PIV system is provided by the MIDR county media library. The collection contains 10 OCRised books dealing with the Pyrenean cultural heritage of the XIXth and XXth century. The books are splitted into paragraphs constituting about ten thousand document units. We have made 12 queries on which 8 deal with only spatial scope whereas the 4 remaining deal with both spatial and thematic scopes. A spatial query could support Absolute Spatial Features (ASF) or Relative Spatial Features (RSF). A thematic and spatial query like "music instruments in Laruns vicinity" supports both ASF/RSF features ("Laruns vicinity") and other non spatial features ("music instruments").

First we carried out scan and OCR processing of the books of the corpora. Then we ran PIV prototype automatic Information Extraction processes. The processing of one book of 200 pages (stages 2, 3 and 4 of Figure 1) takes five minutes. PIV prototype found 9835 candidate SFs in these ten books.

3.3 Evaluation of the Spatial IR Approach

We submitted the eight spatial scope queries to the PIV system and compared the first ranked documents (top 5, 10 and 15) to the hand-craft judgments. The results are given in Table 1. Avg represents the average precision computed over all the used queries and P@5, P@10 and P@15 design precision measures carried out respectively at the top 5, 10 and 15 documents. The last column, Number of responses, represents the total number of retrieved documents (averaged over the queries).

Table 1: PIV and Classical results on spatial queries.

All queries	P@5	P@10	P@15	Number of responses
		A) Spatial approach		
Avg	0.78	0.81	0.73	637
		B) Classical approach		
Avg	0.50	0.43	0.40	252

It can be seen that PIV approach brings 78% accuracy at top 5 and 81% at top 10. When the same queries are applied to the classical full text IR system, the results decrease significantly (Table 1-B). For instance the average precision on the eighth queries at the five top documents (P@5) reaches 78% (PIV) whereas it is only of 50% when using the classical approach. The reason is that in a spatial query like "near Laruns", the classical approach never returns documents dealing with other districts like "Eaux-Bonnes" or "Louvie-Soubiron" which are located in the vicinity of "Laruns". So RSFs extraction from documents and queries also allows increasing the number of retrieved relevant documents: in average 637 document-units are retrieved by the spatial approach for all the queries whereas the classical approach retrieved only 252.

3.4 Evaluation of the Thematic + Spatial IR

We look for the impact of using more general queries containing both spatial and thematic features. As it can be seen in Table 2-A, the results are very decreasing for the PIV approach (only 15% at top 5). A careful analysis of the results shows that some relevant documents are retrieved but they are not ranked at the top. So, PIV system is not suitable for rank-ordering in the case of general (spatial + thematic) queries. Indeed, PIV's IE and IR processes deal only with spatial information.

Table 2: PIV and Classical on thematic + spatial queries.

All queries	P@5	P@10	P@15	Number of responses
A) Spatial approach				
Avg	0.15	0.18	0.18	1154
B) Classical approach				
Avg	0.48	0.39	0.36	331

As in the first case, the same set of queries is submitted to the classical IR system. The results (Table 2-B) are clearly more accurate for the classical approach than those obtained by the PIV system (Table 2-A). For instance, the system brings in average 48% of relevant documents at top 5 and 36% at top15. One can also notices the difference in the number of responses between the two approaches: PIV approach retrieved in average 1154 document-units whereas the classical approach retrieved only 331. This is due to the fact that PIV system processes all spatial features related to the area specified in the query (towns, mountains, etc.), whereas the classical approach seeks for only documents matching the query words.

3.5 Combining Spatial and Classical IR Approaches

The previous results suggest that in the one hand, the spatial PIV approach is suitable to retrieve documents dealing with spatial features but lacks of rank-ordering relevant documents when dealing with non spatial queries. On the other hand, the classical full text approach lacks of exhaustivity when it deals with spatial scope queries but outperforms the PIV approach when the queries deal with thematic features. So, one can think to combine the two approaches in order to take advantage of their effectiveness and reduce their lacks. Moreover, the fact that the document unit corresponds to a paragraph increases the probability that spatial and thematic information occurring in the same unit be semantically related.

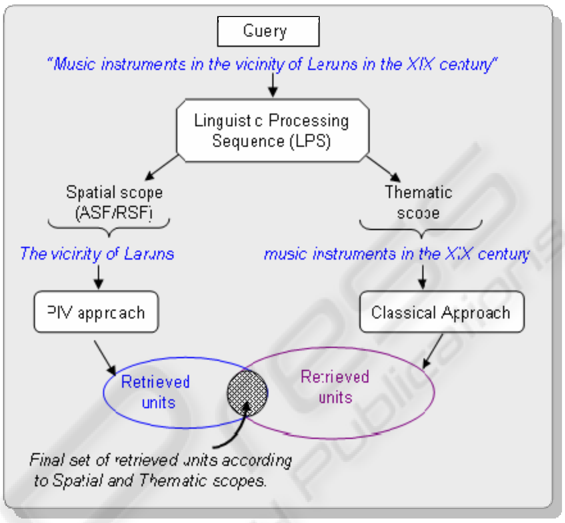

Query
"Music instruments in the vicinity of Laruns in the XIX century"
Linguistic Processing Sequence (LPS)
Spatial scope (ASF/RSF)
Thematic scope
The vicinity of Laruns
music instruments in the XIX century
PIV approach
Classical Approach
Retrieved units
Retrieved units
Final set of retrieved units according to Spatial and Thematic scopes.

Figure 7: Combining Spatial and Classical IR approaches by intersecting the two sets of results.

The idea is to subdivide the query into two sub-queries (as schematized in Figure 7), *the spatial sub-query* and the *thematic one*. The *spatial sub-query* contains named places, or any expression identified by the Linguistic Processing Sequence (LPS) as ASFs or RSFs (cf. §2.3). The thematic *sub-query* contains all the remaining query terms related to any non spatial scope (time, events, etc.) without belonging however to the stoplist. As schematized in Figure 7, "the vicinity of Laruns" and "Music instruments in the XIX century" represents respectively the *spatial sub-query* and the *thematic sub-query* of the query example "Music instruments in the vicinity of Laruns in the XIX century".

Once the two sub-queries are identified, they are submitted to the system supporting the appropriate approach: PIV for the spatial sub-query and Classical for the thematic one. The final result is then built by intersecting the two sets returned by PIV and Classical approaches. The ranking is based on the one obtained by PIV: each ranked document in the PIV result set is added to the final result if it belongs also to the Classical result set.

The detailed results obtained using the previous spatial + thematic queries according to this strategy are given in Table 3. The results confirm the assumption that combining the two approaches will enhance retrieval accuracy by rank-ordering more documents for relevance. For instance at top 5, precision reaches 70% when we combine the two approaches, whereas it was of 48% for the classical approach and only 15% for the spatial approach.

Table 3: Combining PIV with classical approach for the thematic + spatial queries.

All queries	P@5	P@10	P@15	number of responses
A) Combining Spatial + Classical approaches				
Avg	**0.70**	**0.50**	**0.43**	**25.75**

However, one can notice the reduced number of retrieved documents because of the trivial combination used (intersection criteria): for example, fo the query 12, the combined approach retrieves only four documents whereas the Classical approach returns 233 and the PIV one returns 724. This precision improvement causes an important decrease in recall.

So an open area may concern the merging problem of the two sets of results (spatial based approach results and classical full text ones) in order to optimize not only precision at top retrieved documents, but also recall. This may probably be possible by replacing intersection operator by more complex ranking ones.

4 CONCLUSION

Our contribution focuses on restricted corpora such as local cultural heritage collections of documents and is complementary to traditional search methods used in library or documentary management systems. The PIV's Linguistic and semantic processing plus qualitative spatial reasoning support absolute and relative spatial features (ASF/RSF) accurate extraction and retrieval. The PIV prototype validated this approach (Lesbegueries et al., 2006).

A first evaluation scanned the spatial IE process of the PIV prototype (Sallaberry et al., 2007). It led us to extend grammar rules in order to improve the RSF capturing process. We also integrated a new set of spatial resources describing Pyrenean roads, rivers, woods, valleys, mountains, etc.

This paper presents the results of the evaluation of the PIV prototype spatial IR process. A case study involving sample documents and queries given by the MIDR Library of Pau County makes comparisons between the PIV spatial-based prototype and a more classical statistical-based approach. The results show that even-though PIV approach outperforms classical keywords-based approaches in the case of spatial queries. According to these results and those stated in (Vaid et al., 2005), (Martins et al., 2005), such a spatial approach and statistical approaches need to be combined in order to enhance retrieval accuracy in the case of general queries dealing with both spatial and thematic scopes. As the PIV system relies on an architecture of web services, all or part of them might be easily integrated in existing library or documentary management systems.

Such a combined approach's results merging is an actual research point. In fact, PIV's slight IR intersection operator (figure 7) ensures a good precision but a quite poor recall factor. Future works will address integration of spatial and thematic similarity ranking and experiment new merging algorithms using product, maximum similarity, various linear combination functions (Martins et al., 2005).

ACKNOWLEDGEMENTS

Our project is led in partnership with the Greater Pau City Council and the MIDR media library. We want to thank them for providing us with their digital corpus and their support.

REFERENCES

Abolhassani, M., Fuhr, N., Govert; N., 2003. Information Extraction and Automatic Markup for XML documents, *Intelligent Search on {XML} Data*, Springer, vol. 2818, pp. 159–174.

Baeza-Yates, R. A., Ribeiro-Neto., B. A., 1999. *Modern Information Retrieval.* ACM Press / Addison-Wesley.

Borillo, A., 1998. L'espace et son expression en français. *L'essentiel. Ophrys.*

Boughanem, M., Chrisment, C., Tmar, M., 2001. Mercure and MercureFiltre Applied for Web and Filtering Tasks at *TREC-10*. Proceeding of TREC.

Charnois, T., Mathet, Y., Enjalbert, P., Bilhaut, F., 2004. Geographic reference analysis for geographic document querying. *Workshop on the Analysis of Geographic References, Human Language Technology Conference,* NAACL-HLT.

Chen, Y-Y., Suel, T., Markowetz, A., 2006. Efficient query processing in geographic web search engines, *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pp. 277 – 288.

Clementini, E., Sharma, J., and Egenhofer, M., 1994. Modeling topological spatial relations: Strategies for query processing. *Computers and Graphics*.pp. 815-822.

Cohn, A. G., and Hazarika, S. M., 2001. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46(1-2):1-29.

Da Silva, J., Times, V.C., Salgado, A.C., 2006. An Open Source and Web Based Framework for Geographic and Multidimensional Processing. *Advances in Spatial and Image based Information Systems track,* ACM SAC.

Egenhofer, M. J., Franzosa, R.D., 1991. Point-Set Topological Relations. *International Journal for Geographic Information Sytems*, 5(2):161-174.

Egenhofer, M. J., 2002. Toward the semantic geospatial web. In GIS '02: *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, pp. 1–4. ACM Press.

Freeman, J., 1975. The Modelling of Spatial Relations. *Computer Graphics and Image Processing*, 4:156-171.

Gaizauskas, R., Wilks, Y., 1998. Information extraction: Beyond document retrieval. *Journal of Documentation*, 54(1): 70–105.

Gaizauskas, R., 2002. An information extraction perspective on text mining: Tasks, technologies and prototype applications. *Euromap TextMining Seminar*.

Hill, L., 1999. Indirect geospatial referencing through place names in the digital library: Alexandria digital library experience with developing and implementing gazetteers. *62nd Annual Meeting of the American Society for Information Science*, pp. 57-69. Medford, N.J.: ASIS.

Hill, L., 2000. Core elements of digital gazetteers: Place names, categories, and footprints. In ECDL '00: *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, pp. 280–290. Springer-Verlag.

Jones, C.-B., Abdelmoty, A.-I., Finch, D., Fu, G., Vaid, S., 2004. The Spirit Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. *Third International Conference - Geographic Information Science,* Adelphi, Usa, pp. 125 – 139.

Lesbegueries, J., Gaio, M., Loustau, P., and Sallaberry, C., 2006. Geographical information access for non-structured data. *ACM SAC - Advances in Spatial and Image based Information Systems track.*

Lesbegueries, J., Sallaberry, C., and Gaio, M., 2006b. Associating spatial patterns to text-units for summarizing geographic information. *Workshop GIR – SIGIR.*

Malandain, N., Gaio, M., Madelaine, J., 2001. Improving retrieval effectiveness by automatically creating some multiscaled links between text and pictures. In *Proceedings of SPIE, Document Recognition and Retrieval VIII*, volume 4307, pages 89–99.

Martins, B., M. Silva, M-J., and Andrade, L., 2005. Indexing and ranking in Geo-IR systems. *In Proc. of the 2nd Int. Workshop on Geo-IR* (GIR).

Muller, P., 2002. Topological spatio-temporal reasoning and representation. *Computational Intelligence*, pp. 420–450.

Porter, M., 2001. Snowball: A language for stemming algorithms.

http://snowball.tartarus.org/texts/introduction.html

Robertson, S.E., Walker, S., Hancock-Beaulieu, M., Gatford, M., Payne, A., 1995. *Okapi* at TREC-4.

Sallaberry, C., Gaio, M., Lesbegueries, J., and Loustau, P., 2007. A Semantic Approach for Geospatial Information Extraction from Unstructured Documents. In *The Geospatial Web*, Springer. ISBN 1-84628-826-6. http://www.geospatialweb.com/

Sanderson, M. and Kohler, J., 2004. Analyzing geographic queries. In *Proceedings of the Workshop on Geographic Information Retrieval, SIGIR,* www.geo.unizh.ch/~rsp/gir/

Torres, M., 2002. Semantics definition to represent spatial data. *International Workshop -Semantic Processing of Spatial Data* -Geopro.

Vaid, S., Jones, C. B., Joho, H., and Sanderson, M., 2005. Spatio-textual indexing for geographical search on the web. *In Proc. of the 9th Int. Symp. on Spatial and Temporal Databases* (SSTD).

Vandeloise, C., 1986. L'espace en français. Travaux Linguistiques. *Seuil.*

Wildöcher, A., Faurot, E., Bilhaut, F., 2004. Multimodal indexation of contrastive structures in geographical documents. In *RIAO*, pp.555–570.

Widlocher, A., Bilhaut, F., 2005. La plate-forme linguastream : un outil d'exploration linguistique sur corpus. In *Actes de la 12e Conférence Traitement Automatique du Langage Naturel*.

Woodruff, A.G., Plaunt, C., 1994. GIPSY: Automated Geographic Indexing of Text Documents. *Journal of the American Society for Information Science*, 45:9:645-655

Zipf., 1949. Human Behaviour and the Principle of Least Effort. *Addison Wesley*.