Damien Palacio, Guillaume Cabanac, Christian Sallaberry, Gilles Hubert. Measuring Effectiveness of Geographic IR Systems in Digital Libraries: Evaluation Framework and Case Study. In European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2010), Glasgow, Scotland, 06/09/10-10/09/10, Vol. 6273, Mounia Lalmas, Joemon Jose, Andreas Rauber, Fabrizio Sebastiani, Ingo Frommholz (Eds.), Springer, LNCS, p. 340-351, sept. 2010 (Best Paper Award).
http://dx.doi.org/10.1007/978-3-642-15464-5_34

Measuring Effectiveness of Geographic IR Systems in Digital Libraries

Evaluation Framework and Case Study

Damien Palacio[1], Guillaume Cabanac[2],
Christian Sallaberry[1], and Gilles Hubert[2]

[1] Université de Pau et des Pays de l'Adour, LIUPPA ÉA 3000
Avenue de l'Université, BP 1155, F-64013 Pau cedex

[2] Université de Toulouse, IRIT UMR 5505 CNRS
118 route de Narbonne, F-31062 Toulouse cedex 9

Abstract. Common search engines process users' queries (i.e., information needs) by retrieving documents from pre-built term-based indexes. For digital libraries, such approaches are limited regarding particular contexts, such as specialized collections (e.g., cultural heritage collections) or specific retrieval criteria (e.g., multidimensional criteria). In this paper, we consider Information Retrieval systems exploiting geographic dimensions: spatial, temporal, and topical dimensions. Our contribution is twofold as we propose a Geographic Information Retrieval system evaluation framework and test the following hypothesis: combining spatial and temporal dimensions along with the topical dimension improves the effectiveness of Information Retrieval systems.

1 Introduction

Printed literature digitization is currently making significant progress. While some projects only aim to create digital counterparts of physical documents, domain-specific efforts often have more ambitious goals. For example, textual documents are annotated and indexed according to domain-specific models for improving users' experience with document contents [1]. Indeed, Cultural Heritage Libraries generate much digitizing initiatives. The promotion of such collections is then supported by Library Management Systems generally involving full-text Information Retrieval (IR) engines.

In this context, the PIV[3] project, for 'Virtual Itineraries in the Pyrenees Mountains' [2], aims to manage digitized a collection of documents published in the 19th century about the French Pyrenees Mountains. This collection is mainly comprised of newspapers, novels, and travelogues. Local governments foster initiatives aiming at larger dissemination supported by the Web and dedicated IR services. In the meantime, the ratio of geographic queries submitted to usual search engines varies between 12.7% and 18.6% regarding Excite [3], AOL [4],

[3] PIV project is funded by the Pau City Council and the MIDR multimedia library.

and Yahoo [5]. Although query-based common search engines are deemed to deliver accurate results, Kanhabua et al. [6] reported poor precision when it comes to answering geographic queries. As a result, users spend much time skimming the retrieved documents, looking for those that satisfy their information needs. In such a digital library context, it happens that term matching has limitations [7]. For example, the query 'during the 1810's' submitted to a common search engine retrieves only documents containing '1810,' without retrieving '1811,' '1812,' and so on. Similarly, the query 'Paris' retrieves documents containing 'Paris,' but not 'Eiffel Tower,' 'Louvre Museum,' and so on. One way of improving systems accuracy is to include the geographic dimension into the retrieval process. We consider the usual acceptance that Geographic Information gathers three dimensions, namely spatial, temporal and, topical [8]. A typical illustration of this is: 'Fortified towns in South London suburbs during the 13th century.'

Following up recent work on Digital Libraries [7], the main goal of PIV project is to help users finding accurate information inside books. We intend to overcome usual IR Systems (IRSs) limitations regarding geographic information management. Thus, we designed three process chains for indexing spatial [2], temporal [9] and topical [10] information. These allow the retrieval of document units along with associated relevance scores. This work involves various domains, such as Natural Language Processing (NLP), Geographic Information Systems (GISs), Information Retrieval (IR), and Geographic Information Retrieval (GIR). More specifically, this paper tackles both design and experiment of an evaluation framework dedicated to GIR systems. This framework *i*) addresses the evaluation of spatial, temporal and topical IRSs, as well as the evaluation of any combination of these; *ii*) designs a test collection suited to GIR context, which may lead us to implement an original GeoCLEF [11] -like track.

The paper is organized as follows. Existing evaluation frameworks dedicated to GIR systems are presented in Sect. 2. In Sect. 3, we propose an evaluation framework addressing GIR systems. Section 4 develops a case study: it presents and evaluates PIV core components—spatial, temporal, and topical IRSs—and their combination. Our hypothesis is: combining these three dimensions is more effective than any of the underlying IRSs. This is tested through an experiment complying with the proposed evaluation framework. In Sect. 5, we review the literature related to GIR systems; these may benefit from our framework. Finally, Sect. 6 concludes the paper and outlines research directions.

2 Suitability of Existing Evaluation Frameworks for GIR

IR has a long tradition of experimentation, especially regarding TREC [12] program dedicated to topical IR evaluation. Moreover, TEMPEVAL [13] evaluation framework is concerned with the temporal dimension. Building on both of these initiatives, Bucher et al. [14] proposed to evaluate two dimensions at the same time: spatial and topical dimensions. This proposal was realized in GeoCLEF [11] task of CLEF program [15]. It notably allowed Perea-Ortega et al. [16] to evaluate the effectiveness of classical topical IRSs, such as Terrier [17].

To the best of our knowledge, GIR contributions (reviewed in Sect. 5) were mostly evaluated according to efficiency (e.g, index size and retrieval performance in time). However, it may be worth complementing such quantitative figures with effectiveness evaluation. Moreover, no work to date considered evaluating the three dimensions altogether. Consequently, it is not possible to compare search engines handling these features yet. That is the reason why we propose in the next section an evaluation framework dedicated to GIR.

3 Design of a Framework for GIR Evaluation

The proposed evaluation framework builds on existing state-of-the-art methodologies (especially related to TREC and GEOCLEF), and integrates the lacking specificities regarding geographic information. Section 3.1 details the design of a test collection covering the three geographic dimensions; then Sect. 3.2 reports the analysis of PIV GIR system, enabling us to assess its effectiveness.

3.1 Test Collection Supporting GIR Evaluation

In the IR literature, especially at TREC [18], a test collection is comprised of the three following components:

1. A set of n *topics* representing users' information needs. Each topic is at least provided with a title (a keyword-based query), a description (usually a sentence in ordinary language), and a narrative (a detailed explanation of expected information as well as criteria for judging a document as relevant or non-relevant). While a minimum of 25 topics are required for conducting sound statistical analyses [18], note that 50 topics is standard at TREC.
2. The *corpus* comprising numerous documents, some of which are relevant for the proposed topics. A regular TREC corpus for classical ad hoc task represents from 800,000 to 1 million documents [18].
3. The *qrels* (i.e., query relevance judgments) associating each topic with the documents that an individual would expect to retrieve, i.e., a set of relevant documents. Since the corpus is too huge to be extensively considered looking for relevant documents, IR evaluation frameworks rely on the 'pooling' technique, especially at TREC. For each topic t, a document pool is created from the top 100 documents retrieved by the participants' IRSs, duplicates being removed. It is hypothesized that resorting to multiple and diverse IRSs leads to finding most of the relevant documents belonging to the corpus. Finally, an assessor skims through each document for evaluating whether it matches the information need corresponding to topic t or not: the document is then qualified as *relevant* or *non-relevant*.

Such test collections were operated for several evaluation frameworks, especially at TREC and GEOCLEF. Notice that they do not cover all the three dimensions of geographic information. This motivates our work, as we propose to customize the design of test collections in order to enable GIR evaluation by providing:

1. *Topics* covering part or the totality of the three dimensions. For instance, a topic may be titled 'Potato Famine in Southern Eire after mid-19th century' and its narrative may be 'Relevant documents mention scarcity of food and its consequences in Southern Ireland after 1849.'
2. A *corpus* covering the three dimensions: documents present not only the usual topical items but also additional spatial and temporal items.
3. *Qrels* associated with each dimension, resulting from the judgment of relevance between documents and the three dimensions (topical, spatial, and temporal). The co-occurrence of these three dimensions in a given document is not enough for deducing its relevance with respect to the query. Let us consider a document citing 'Dublin City' as the protagonist's place of birth. Although spatially relevant, it does not match the query 'Pubs in Dublin.' This subtlety requires the assessment of the global match between the query and the document. Not to overwhelm assessors, we opted for a per dimension binary judgment: a document is either relevant or non-relevant to the considered query and dimension. This rationale is akin to Bucher and colleagues' conclusions about gradual judgments for each dimension, which were judged as 'unnecessary cumbersome' [14]. Finally, considering the three per-dimension binary judgments, as well as the aforementioned global binary judgment, we compute the document relevance value $v \in \{0, 1, 2, 3, 4\}$. This both represents the number of satisfied dimensions, and global relevance. No assumption was made regarding the relative importance of dimensions; they were equitably considered.
4. *Geographic resources* georeferencing spatial entities that occur in the corpus.

The experimental protocol detailed in the next section intends to measure the effectiveness of GIRSs. They are evaluated on the basis of the *runs* they provided, i.e., the retrieved document list per topic.

3.2 Protocol for GIRS Comparative Evaluation

The task under evaluation is called 'ad hoc' at TREC: an IRS addresses a query by providing a document list ranked by decreasing score. Indeed, the evaluation framework allows effectiveness evaluation for the following IRSs: monodimensional (topical To, spatial S, and temporal Te), bidimensional (To+S, To+Te, and S+Te allowing the measurement of effectiveness improvement according to each missing dimension), and GIRS combining the three dimensions (To+S+Te).

For a given topic, each IRS provides a list of pairs (d, s) representing the score s of each retrieved document d. Usually, effectiveness of an IRS is evaluated with respect to *Average Precision* (*AP*) measure for each topic, and *Mean Average Precision* (*MAP*) overall. These require binary qrels [19, ch. 8]. In the protocol that we propose, however, qrels are gradual in order to represent the three dimensions of geographic information. These two measures are not suitable indeed. We thus used *Normalized Discounted Cumulative Gain* (*NDCG*) [20] relying on gradual relevance judgments; it was notably used at TREC-9 for the

Web task [18]. *NDCG* implements two principles. On the one hand, highly relevant documents ($v \to 4$ in our case) are more valuable than marginally ($v \to 1$) relevant documents. On the other hand, a document is all the less valuable that it is ranked low in the result list, because it is rather unlikely that the user reaches this document. Following the example of TREC, we propose two granularity levels for evaluating an IRS: *i*) topic level is represented by *NDCG* while *ii*) the overall level is computed by *MANDCG*: the mean average of the n *NDCG* values, giving the overall effectiveness of an IRS.

For the overall level, the observed differences $\langle m_i^1 - m_j^1, \ldots, m_i^n - m_j^n \rangle$ between two systems for the n topics are reported in per cent (of increase or decrease). We denote m_s^t as the value of measure m achieved by system s for topic t. Statistical test significance of the paired observed differences are also reported with respect to significance p-values resulting from Student's paired two-tailed t-test. Although this test theoretically requires a normal data distribution, Hull [21] states that it is robust to violations of this condition. In practical terms, when $p < \alpha$ where $\alpha = 0.05$ the difference between the tested samples is statistically significant; the smaller p-value, the more significant the difference [21].

4 Case study: PIV GIR System Evaluation

We experiment PIV prototype with our evaluation framework in order to validate the aforementioned hypothesis: combining the three geographic dimensions improves retrieval effectiveness. Therefore, Sect. 4.1 describes this prototype and Sect. 4.2 reports results of its evaluation.

4.1 PIV GIR System

Indexing: Spatial, Temporal and Topical Process Chains. As proposed by Clough et al. [22], we process each of the three geographic dimensions independently. This can be achieved by building several indexes, one per dimension, as advised in [23]. In this way, one can restrain the search on one criterion and easily manage the indexes (e.g., allowing document addition to the corpus). So, our approach processes indexes independently and combines them later on for supporting multidimensional IR. It contributes to GIR field as defined by Jones and Purves [24], as well as GIR in Digital Libraries as defined by Larson [25].

PIV implements three indexing process chains dedicated to textual document processing. Each process chain builds an index. Spatial and temporal process chains are supported by dedicated NLP services. They provide spatial (SF) and temporal/calendar (CF) feature extraction from textual documents and their interpretation: 'River Thames' is annotated as an absolute SF whereas 'North of the River Thames' is annotated as a relative SF—spatial orientation relation [2]. In the same way, 'Spring 1840' is annotated as an absolute CF whereas 'around Spring 1840' is annotated as a relative CF—temporal adjacency relation [9]. So, spatial and temporal indexes result from various stages. The first stage consists in a syntactico-semantic processing sequence [2]: it addresses SF and CF extraction.

This stage is supported by the LinguaStream platform [26]. It is mainly comprised of lexical analysis, morpho-syntactic analysis, and syntactico-semantic analysis relying on DCG (Definite Clause Grammar) rules intending to associate one type and one semantics to any extracted SF and CF. The second stage aims at SF and CF interpretation. This uses the symbolic representation of any SF and CF, and operates specific algorithms to calculate approximated numeric representations: spatial geometries are computed for SF [2] and time intervals are computed for CF [9]. Finally, the third stage standardizes resulting indexes. It consists in a spatial, temporal, and topical tiling process that computes the frequency of these spatial, temporal, and topical tiles within texts and weights them [27]. The resulting indexes allow the implementation of state-of-the-art models for relevance score computation based on such spatial, temporal or topical tiles given their occurrence frequencies within documents. This is discussed in the next section, which details such original strategies.

Retrieval: Combination of Result Lists. These indexing strategies may be associated with discrete or continuous scores depending on the overlap ratio between the SF and the tile (the CF and the tile). This enables us to weigh a tile accordingly. We conducted experiments involving IR weighting schemes (TF, TF·IDF, OkapiBM25) along with discrete and continuous frequency computations. TF formula associated with continuous frequency gave the best results in our context [27]. That is the reason why we evaluate combinations of PIV's spatial and temporal retrieval results hereafter. However, as PIV topical process chain is not fully automated yet, we restrain PIV's topical component to the state-of-the-art Terrier full-text IRS.

Each monodimensional IRS is independent: it builds and queries its own index. PIV is supported by these three different monodimensional (source) IRS. Their results are combined in order to constitute a single result list l. Now, Fox and Shaw [28] introduced the CombMNZ combination operator in the IR field; the resulting combined list l gathers together distinct documents retrieved by the source IRSs. Therefore, the similarity s of a document d in l is computed by adding the similarities of d extracted from source IRSs. This sum is balanced by the number of source IRSs that retrieved d. As a result, for any query q, the higher d is ranked in the result lists of the source IRSs (high similarity s between d and q), the more relevant in l it is (i.e., ranked in the top of l). CombMNZ may be compared to a burden of proof, gathering pieces of evidence: documents retrieved by several source IRSs are so many clues enforcing their presumption of relevance. We validated this principle in a quite different context involving combination of the topical and the semantic dimensions [29].

In addition, Lee [30] compared CombMNZ with other operators on TREC test collections, and demonstrated its effectiveness. That is the reason why we experimented with this operator for combining monodimensional IRS results. As every similarity value s computed by source IRSs may belong to a distinct numeric domain, we normalized them within $[0, 1]$ according to: $\text{normalized\_similarity} = \frac{\text{unnormalized\_similarity} - \text{minimum\_similarity}}{\text{maximum\_similarity} - \text{minimum\_similarity}}$ [30]. So, for query $q = 8$, Fig. 1 (a–c) il-

lustrates results retrieved by three IRSs: each result list is comprised of (d_i, s) pairs where d_i is a document and s is the computed similarity between q and d_i. Combination of these IRSs results is detailed in Fig. 1 (d). It shows CombMNZ similarity values and the corresponding computation details. These similarity values are based on the normalized values of IRS sources, cf. Fig. 1 (a–c).

(a) Topical IRS

q	d	s
8	d_4	14.5
8	d_3	12
8	d_7	8.7
8	d_1	0.5

(b) Spatial IRS

q	d	s
8	d_8	150
8	d_1	120
8	d_4	80
8	d_9	−10
8	d_2	−30

(c) Temporal IRS

q	d	s
8	d_8	1
8	d_4	0.7
8	d_9	0.5
8	d_1	0.5
8	d_2	0.5

↘ ↓ ↙

(d) Geographic IRS resulting from merged source result lists

q	d	Similarity s resulting from CombMNZ
8	d_4	$6.0333 = 3 \cdot \left(\frac{14.5-0.5}{14.5-0.5} + \frac{80+30}{150+30} + \frac{0.7-0.5}{1-0.5}\right)$
8	d_8	$4.0000 = 2 \cdot \left(\frac{150+30}{150+30} + \frac{1-0.5}{1-0.5}\right)$
8	d_1	$2.5000 = 2 \cdot \left(\frac{0.5-0.5}{14.5-0.5} + \frac{120+30}{150+30}\right)$
8	d_3	$0.8214 = 1 \cdot \left(\frac{12-0.5}{14.5-0.5}\right)$
8	d_7	$0.5857 = 1 \cdot \left(\frac{8.7-0.5}{14.5-0.5}\right)$
8	d_9	$0.2222 = 2 \cdot \left(\frac{-10+30}{150+30} + \frac{0.5-0.5}{1-0.5}\right)$
8	d_2	$0.0000 = 2 \cdot \left(\frac{-30+30}{150+30} + \frac{0.5-0.5}{1-0.5}\right)$

Fig. 1. Illustration of result lists combination with CombMNZ [28]

The example in Fig. 1 illustrates how the score of a d_i takes into account two factors. The more often a source IRS retrieves document d_i, the higher its score s is. Additionally, the higher an IRS ranks d_i in the associated result list, the higher its score s is. In particular, document d_4 illustrates this principle.

Notice that the focus of this paper is on GIRS evaluation. As a result, and due to space limitation, we do not explain PIV components in greater detail. Nevertheless, the interested reader may refer to [2, 9, 27] in this respect.

4.2 Evaluation Framework Use for Evaluating the PIV System

We applied the evaluation framework presented in Sect. 3 for evaluating the PIV system. We consider in this section the constituted test collection, the comparative analyses that we carried out, and their limitations.

Design of the MIDR_2010 Test Collection. The MIDR_2010 test collection is comprised of the four components identified in Sect. 3.1. First, the *corpus* collects 5,645 paragraphs extracted from 11 books published between the 18th and 20th centuries, and belonging to Aquitaine Regional Library. They were scanned

and processed with OCR software. A document d, as retrieved by the IRS, is one of these paragraphs; it is considered as the best entry point in its associated book. Second, 31 *topics* covering the three dimensions of geographic information were constituted. Third, *qrels* were obtained by querying three IRS—a topical IRS based on PL2 IR model (built-in Terrier configuration), a spatial IRS and a temporal IRS—with the 'title' part of the topics. Lastly, the *geographic resources* corresponding to the corpus are provided by the French National Geographic Institute (BD NYME ® database). For each topic, the results retrieved by the IRSs were considered for setting up the pool. It was then assessed according to binary judgments for each dimension, and for the global judgment. These four judgments were aggregated in a single gradual value, as presented in Sect. 3.1.

Comparative Analysis of IRS Effectiveness. In Tab. 1, we report the observed comparisons between various IRSs and two baselines identified in [16]: To$^+$ is a strong baseline corresponding to OkapiBM25 model; To$^-$ is a weaker baseline corresponding to TF·IDF model. In addition, S denotes the spatial IRS, and Te denotes the temporal IRS. The reported results show effectiveness of search engines according to the 31 tested topics. Overall, the two baselines showed similar effectiveness. Contrary to the observations presented in [16], TF·IDF achieves better performance than OkapiBM25 in our experiment. This difference may be due to the fact that the MIDR_2010 test collection is comprised of document paragraphs similar in length, as opposed to plain documents with variable lengths, for which OkapiBM25 is known to achieve best performance.

Table 1. IRS effectiveness w.r.t. topical baselines. The * symbol († symbol) denotes a significant difference compared with baseline To$^-$(baseline To$^+$)

Combination of N IRSs	Monodimensional IRS					Gain (%)	
	To$^-$	To$^+$	S	Te	*MANDCG*	To$^-$	To$^+$
1	✓				0.4726	0.0	0.1
		✓			0.4721	−0.1	0.0
			✓		0.4574	−3.2	−3.1
				✓	0.4836	2.3	2.4
2	✓	✓			0.4722	−0.1	0.0
	✓		✓		0.6162*†	30.4	30.5
	✓			✓	0.7017*†	48.5	48.6
		✓	✓		0.6165*†	30.4	30.6
		✓		✓	0.7017*†	48.5	48.6
			✓	✓	0.6993*†	48.0	48.1
3	✓	✓	✓		0.6104*†	29.2	29.3
	✓	✓		✓	0.6842*†	44.8	44.9
	✓		✓	✓	0.7852*†	66.1	66.3
		✓	✓	✓	**0.7859***†	66.3	66.5
4	✓	✓	✓	✓	0.7578*†	60.3	60.5

Regarding monodimensional IRSs they all achieve similar performance; best effectiveness (0.4836) is reached by the temporal IRS. In addition, combining at least two heterogeneous dimensions yields better performance. Notice that the associated improvement is statistically significant regarding the two baselines, Te+To$^+$ and Te+To$^-$ being the most effective combinations (0.7017). However, the combination Te+S is similar in effectiveness (0.6993). An explanation for this may involve absolute spatial entities (e.g., 'Paris'), which are easily retrieved by a topical IRS (exact match). However, only a spatial IRS can properly process more complex queries involving relative spatial entities (e.g., 'Eastern Paris').

Combining the three dimensions (0.7859) yields better results (+12.0%) than the best bidimensional combination (0.7017), which is statistically significant ($p = 0.000$). Adding to these dimensions a topical IRS (To$^+$ or To$^-$) does not result in more improvement (0.7578). The resulting topical reinforcement may lessen the complementary information contributed by the two other dimensions. In conclusion, combining the three dimensions provides the best performance (0.7859). The 66.3% improvement regarding To$^-$ validates the hypothesis formulated in this paper: the combination of the three geographic information dimensions yields better performance than considering only the topical dimension.

Limitations of the Current Evaluation. The experiment reported in this paper presents at least two limitations. On the one hand, comprising 5,645 paragraphs for a 3.7 Mb total size, the MIDR_2010 test collection is very limited in size compared with TREC collections. On the other hand, the experiment was completed with 31 topics; this represents six more topics than the minimum number of topics to use for conducting proper statistical analysis. We keep on judging documents manually in order to provide more topics.

Despite these limitations, the evaluation framework proposed in this paper is appropriate for experimenting with the various proposals found in the GIR domain. The next section briefly introduces a representative sample of this work.

5 GIR Related Work

Work related to GIR includes the following prominent five projects. GIPSY [31], for 'Georeferenced Information Processing System,' proposes a method for indexing textual documents; it is based on the aggregation of the footprints corresponding to spatial entities. This aggregation is used to find the most representative geographic areas in order to index a document. GeoSem [26], for 'Geographic Semantic,' is dedicated to document (texts, maps, charts) geographic information semantics processing. SPIRIT [32], for 'Spatially-Aware Information Retrieval on the Internet,' aims to find Web pages that refer to places or geographic areas specified in a query. STEWARD [33], for 'Spatio-Textual Extraction on the Web Aiding Retrieval of Documents', performs extraction, retrieval, and geographic area visualization for unstructured texts. CITER [34] for 'Creation of a European History Textbook Repository' offers history textbook retrieval according to several dimensions. Finally, DIGMAP [35], for 'Discovering our Past

World with Digitised Maps,' is dedicated to cultural and scientific heritage promotion. These geographic information indexing and retrieval systems handle in priority the spatial criteria (SF, cf. Sect. 4.1). They all use pre-built monodimensional indexes and propose similar approaches to merge result lists. They apply filtering-like approaches: for instance, STEWARD retrieves topical relevant document units first, and goes on filtering out these results according to the spatial dimension. This is quite different from the CombMNZ [28] combination-based approach that we introduced in this paper.

6 Conclusion and Future Work

We considered geographic IRSs handling spatial, temporal, and topical dimensions. However, common search engines show limitations in such contexts. Consequently, our contribution is twofold: we propose an evaluation framework, and use it for validating our hypothesis: combining the three dimensions improves the accuracy of retrieval results. Applying this framework on an appropriate test collection showed an improvement of 66.3% over a topical baseline. Moreover, this performance gain is statistically significant. These good results give an empirical validation of our proposals experimented with the PIV GIR system [2]. In addition, this evaluation framework is not restricted to the three mentioned dimensions: it can also integrate other dimensions, such as confidence in the information, and its freshness [36].

In addition to experimenting with a larger test collection, we are now intending to propose and experiment alternate IRSs combination approaches. We are especially interested in constraint-based combination methods—involving concepts of requirement and preference—based on a linear approach [37], or related to fuzzy OWA [38] -based approaches. Having demonstrated the feasibility of GIR evaluation in this paper, we plan to propose this framework in a GeoCLEF-like track. Its main originality is to provide documents, and allow the evaluation of IRSs according to the three dimensions of geographic information (i.e., topical, spatial and temporal). For this purpose, the constituted MIDR_2010 test collection is available on PIV project website[4].

References

1. Sautter, G., Böhm, K., Padberg, F., Tichy, W.F.: Empirical Evaluation of Semi-automated XML Annotation of Text Documents with the GoldenGATE Editor. In: ECDL'07: 11th European Conference on Digital Libraries. Volume 4675 of LNCS., Springer (2007) 357–367
2. Gaio, M., Sallaberry, C., Etcheverry, P., Marquesuzaa, C., Lesbegueries, J.: A global process to access documents' contents from a geographical point of view. J. Vis. Lang. Comput. **19**(1) (2008) 3–23
3. Sanderson, M., Kohler, J.: Analyzing Geographic Queries. In: SIGIR-GIR'04: Workshop on Geographic Information Retrieval at SIGIR. (2004)

[4] http://t2i.univ-pau.fr/MIDR/

4. Gan, Q., Attenberg, J., Markowetz, A., Suel, T.: Analysis of geographic queries in a search engine log. In: LocWeb'08: 1st Int. workshop on Location and the web, New York, NY, USA, ACM (2008) 49–56
5. Jones, R., Zhang, W.V., Rey, B., Jhala, P., Stipp, E.: Geographic intention and modification in web search. Int. J. Geogr. Inf. Sci. **22**(3) (2008) 229–246
6. Kanhabua, N., Nørvåg, K.: Improving Temporal Language Models for Determining Time of Non-timestamped Documents. In: ECDL'08: 12th European conference on Research and Advanced Technology for Digital Libraries, Berlin, Heidelberg, Springer (2008) 358–370
7. Liesaputra, V., Witten, I.H., Bainbridge, D.: Searching in a Book. In: ECDL'09: 13th European Conference on Digital Libraries. Volume 5714 of LNCS., Springer (2009) 442–446
8. Usery, E.L.: A feature-based geographic information system model. Photogramm. Eng. Rem. Sens. **62**(7) (1996) 833–838
9. Le Parc-Lacayrelle, A., Gaio, M., Sallaberry, C.: La composante temps dans l'information géographique textuelle. Document Numérique **10**(2) (2007) 129–148
10. Sallaberry, C., Baziz, M., Lesbegueries, J., Gaio, M.: Towards an IE and IR System Dealing with Spatial Information in Digital Libraries – Evaluation Case Study. In: ICEIS'07: 9th Int. Conference on Enterprise Information Systems. (2007) 190–197
11. Gey, F.C., Larson, R.R., Sanderson, M., Joho, H., Clough, P., Petras, V.: GeoCLEF'05: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In: CLEF'05: 6th workshop on Cross-Language Evalution Forum. Volume 4022 of LNCS., Springer (2006) 908–919
12. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge, MA, USA (2005)
13. Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., Pustejovsky, J.: The TempEval challenge: identifying temporal relations in text. Lang. Resour. Eval. **43**(2) (2009) 161–179
14. Bucher, B., Clough, P., Joho, H., Purves, R., Syed, A.K.: Geographic IR Systems: Requirements and Evaluation. In: ICC'05: 22nd Int. Cartographic Conference, Global Congressos (2005) CDROM.
15. Peters, C.: Introduction. In: CLEF'01: 1st Workshop Cross-Language Information Retrieval and Evaluation. Volume 2069 of LNCS., Springer (2001) 1–6
16. Perea-Ortega, J.M., García-Cumbreras, M.A., García-Vega, M., Ureña-López, L.A.: Comparing Several Textual Information Retrieval Systems for the Geographical Information Retrieval Task. In: NLDB'08: 13th Int. conference on Natural Language and Information Systems, Berlin, Heidelberg, Springer (2008) 142–147
17. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier Information Retrieval Platform. In: ECIR'05: 27th European Conference on IR Research. Volume 3408 of LNCS., Springer (2005) 517–519
18. Harman, D.K.: The TREC Test Collections. [12] chapter 2 21–53
19. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (July 2008)
20. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. **20**(4) (2002) 422–446
21. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In: SIGIR'93: 16th annual Int. SIGIR conference, New York, NY, USA, ACM Press (1993) 329–338
22. Clough, P., Joho, H., Purves, R.: Judging the Spatial Relevance of Documents for GIR. In: ECIR'06: 28th European Conference on IR Research. Volume 3936 of LNCS., Springer (2006) 548–552

23. Martins, B., Silva, M.J., Andrade, L.: Indexing and ranking in Geo-IR systems. In: GIR '05: workshop on Geographic information retrieval, New York, NY, USA, ACM (2005) 31–34
24. Jones, C.B., Purves, R.: GIR'05 2005 ACM workshop on geographical information retrieval. SIGIR Forum **40**(1) (2006) 34–37
25. Larson, R.R.: Geographic Information Retrieval and Digital Libraries. In: ECDL'09: 13th European Conference on Digital Libraries. Volume 5714 of LNCS., Springer (2009) 461–464
26. Bilhaut, F., Charnois, T., Enjalbert, P., Mathet, Y.: Geographic reference analysis for geographic document querying. In: HLT-NAACL'03: workshop on Analysis of geographic references, Morristown, NJ, USA, ACL (2003) 55–62
27. Palacio, D., Sallaberry, C., Gaio, M.: Normalizing Spatial Information to Improve Geographical Information Indexing and Retrieval in Digital Libraries. In: ISGIS'10: Joint Int. Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science proceedings. (2010) to appear.
28. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In Harman, D.K., ed.: TREC-1: 1st Text REtrieval Conference, Gaithersburg, MD, USA, NIST (February 1993) 243–252
29. Hubert, G., Mothe, J.: An adaptable search engine for multimodal information retrieval. J. Am. Soc. Inf. Sci. Technol. **60**(8) (2009) 1625–1634
30. Lee, J.H.: Analyses of Multiple Evidence Combination. In: SIGIR'97: 20th annual Int. SIGIR conference, New York, NY, USA, ACM Press (1997) 267–276
31. Woodruff, A.G., Plaunt, C.: Gipsy: automated geographic indexing of text documents. J. Am. Soc. Inf. Sci. **45**(9) (1994) 645–655
32. Vaid, S., Jones, C.B., Joho, H., Sanderson, M.: Spatio-textual Indexing for Geographical Search on the Web. In: SSTD'05: 9th Int. Symposium on Spatial and Temporal Databases. Volume 3633 of LNCS., Springer (2005) 218–235
33. Lieberman, M.D., Samet, H., Sankaranarayanan, J., Sperling, J.: STEWARD: Architecture of a Spatio-Textual Search Engine. In: GIS'07: 15th annual ACM Int. symposium on Advances in geographic information systems, New York, NY, USA, ACM (2007) 1–8
34. Pfoser, D., Efentakis, A., Hadzilacos, T., Karagiorgou, S., Vasiliou, G.: Providing Universal Access to History Textbooks: A Modified GIS Case. In: W2GIS'09: 9th Int. Symposium on Web and Wireless Geographical Information Systems. Volume 5886 of LNCS. (2009) 87–102
35. Manguinhas, H., Martins, B., Borbinha, J., Siabato, W.: The DIGMAP Geo-Temporal Web Gazetteer Gervice. e-Perimetron: Int. Web J. Sci. Technol. Affined Hist. Cartogr. Maps **4**(1) (2009) 9–24
36. Costa Pereira, C., Dragoni, M., Pasi, G.: Multidimensional relevance: A new aggregation criterion. In: ECIR'09: 31th European Conference on IR Research on Advances in Information Retrieval, Berlin, Heidelberg, Springer (2009) 264–275
37. Farah, M., Vanderpooten, D.: An outranking approach for information retrieval. Inf. Retr. **11**(4) (2008) 315–334
38. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Trans. Syst. Man Cybern. **18**(1) (1988) 183–190