Preprint of the article published in Int J Digit Libr, 11(2):91-109, 2010
DOI 10.1007/s00799-011-0070-z

On the evaluation of Geographic Information Retrieval systems

Evaluation framework and case study

Damien Palacio · Guillaume Cabanac · Christian Sallaberry · Gilles Hubert

Received: December 14, 2010 / Revised: June 7, 2011 / Accepted: August 19, 2011 / Published online: September 28, 2011

Abstract Search engines for Digital Libraries allow users to retrieve documents according to their contents. They process documents without differentiating the manifold aspects of information. Spatial and temporal dimensions are particularly dismissed. These dimensions are, however, of great interest for users of search engines targeting either the Web or specialized Digital Libraries. Recent studies reported that nearly 20% queries convey spatial and temporal information in addition to topical information. These three dimensions were referred to as parts of 'geographic information.' In the literature, search engines handling those dimensions are called 'Geographic Information Retrieval (GIR) systems.' Although several initiatives for evaluating GIR systems were undertaken, none was concerned with evaluating these three dimensions altogether. In this article, we address this issue by designing an evaluation framework, whose usefulness is highlighted through a case study involving a test collection and a GIR system. This framework allowed the comparison of our GIR system to state-of-the-art topical approaches. We also performed experiments for measuring performance improvement stemming from each dimension or their combination. We show that combining the three dimensions yields improvement in effectiveness (+73.9%) over a common topical baseline. Moreover, rather than conveying redundancy, the three dimensions complement each other.

Keywords Geographic Information Retrieval · Effectiveness Measurement · Evaluation Framework · Case Study

1 Introduction

Digital Libraries (DL) is an interdisciplinary and active research field, whose results have been implemented in several worldwide initiatives. One of its latest prominent application is the Europeana project, a cross-domain cultural heritage portal funded by the European Commission [1]. Such projects aim at collecting patrimonial documents in order to keep and promote them. Those documents are all the more valuable as they cover various topics, various temporal periods, and various locations. They are made available through search engines: Information Retrieval (IR) systems allowing people to retrieve documents matching their information needs. Such systems usually match query terms with document contents and metadata. This process may seem sufficient enough for simple queries (e.g., *'Potato famine'*). However, it may turn out to be inaccurate for more complex or elaborated information needs (e.g., *'Potato famine in Ireland after the mid-19th century'*). In such queries, several dimensions are considered, indeed. In addition to the topical dimension (*Potato famine*), spatial (*Ireland*), and temporal (*after the mid-19th century*) dimensions are present. In the literature, information characterized by these three dimensions is known as 'geographic information' [2]. The proportion of geographic queries submitted to common search engines varies between 12.7% and 22.7%, as reported in Tab. 1.

This paper is a substantially revised and extended version of a paper (with the title 'Measuring Effectiveness of Geographic IR Systems in Digital Libraries: Evaluation Framework and Case Study') originally appeared in the Proceedings of the 14th European Conference on Digital Libraries (ECDL 2010).

Electronic Supplementary Material The online version of this article (doi:10.1007/s00799-011-0070-z) contains supplementary material, which is available to authorized users.

D. Palacio · C. Sallaberry
Université de Pau et des Pays de l'Adour, LIUPPA ÉA 3000
Avenue de l'Université, BP 1155, F-64013 Pau cedex
E-mail: damien.palacio@univ-pau.fr, christian.sallaberry@univ-pau.fr

G. Cabanac · G. Hubert
Université de Toulouse, IRIT UMR 5505 CNRS
118 route de Narbonne, F-31062 Toulouse cedex 9
E-mail: guillaume.cabanac@irit.fr, gilles.hubert@irit.fr

Table 1 Proportion of geographic queries submitted to common search engines

Year	Reference	Search Engine	Proportion (%)
2004	Sanderson & Kohler [3]	Excite	18.6
2005	Asadi et al. [4]	MetaCrawler	22.7
2005	Souza et al. [5]	TodoBR	14.1
2008	Gan et al. [6]	AOL	13.0
2008	Jones et al. [7]	Yahoo!	12.7

Since common search engines consider only the topical dimension of documents, they tend to yield partial results when missing to address two dimensions, namely spatial and temporal dimensions. Therefore, the quality of results returned to the user is underachieved. As a mean to tackle this retrieval quality issue, researchers in the Geographic Information Retrieval (GIR) field built on various techniques stemming from multiple domains, such as Natural Language Processing (NLP), Geographic Information Systems (GISs), as well as Information Retrieval. Several GIR systems were proposed in the literature; they process the topical dimension of documents along with other dimensions (i.e., spatial, temporal), as a way to improve search engine effectiveness. Yet most of them handle only two dimensions out of the three (i.e., topical and temporal dimensions, or topical and spatial dimensions). Moreover, systems were mostly evaluated according to their efficiency (i.e., performance measured in time and space units), and initiatives of effectiveness evaluation (i.e., relevance of search results) do not cover the three dimensions. As a result, there is currently no means of evaluating the full potential of GIR systems.

In this article, we present the following contributions:

1. The design of an *evaluation framework* dedicated to GIR systems. This evaluates the effectiveness of systems according to the three dimensions of geographic information: spatial, temporal, and topical dimensions. In addition, it allows to compare the performances of any GIR system to common topical search engines, as well as to investigate the complementarity of the three dimensions of geographic information.
2. The experiment of this evaluation framework on a *case study* for highlighting its usefulness. This involves:
   - The design of the MIDR_2010 test collection (Online Resource 1) suited to GIR evaluation, as it covers the three aforementioned dimensions.
   - The PIV ('Virtual Itineraries in the Pyrenees mountains') GIR system on the basis of [8]. PIV is a search engine that processes queries on the three dimensions of geographic information. It relies on three process flows for indexing documents according to their spatial [8], temporal [9] and topical [10] contents. On the retrieval stage, the results from each index supporting a dimension are combined to form a single list of documents returned to the user.

The article is organized as follows. In Sect. 2, we review the literature related to GIR systems. In Sect. 3 we present the PIV GIR system that we designed and detail its core components — spatial, temporal, and topical IR systems — and the way they are combined together. Our hypothesis is: combining these three dimensions is more effective than using any of these alone. In Sect. 4, we review the literature related to the evaluation of GIR systems. In Sect. 5, we propose an evaluation framework addressing GIR systems. As a case study, we implemented and experimented the proposed evaluation framework to evaluate our PIV system in Sect. 6. In Sect. 7, we discuss our contributions, their evaluation, and limitations. Finally, we conclude in Sect. 8 and outline research directions. Notice that the abbreviations used in this article are summarized in the appendix (Tab. 9).

2 State of the art of GIR

In this section, we briefly present concepts related to GIR, as well as prominent GIR systems from the literature.

2.1 GIR concepts

In the context of textual document repositories, Natural Language Processing tasks contribute to Information Extraction (IE) and Retrieval (IR) processes (Fig. 1). IE consists in extracting information from unstructured documents, represents it with appropriate descriptors and organizes it into structured indexes for supporting IR. IR consists in matching user needs with indexed document descriptors. Finally, Information Visualization (IV) displays results of an IR process and supports browsing scenarios, as well as feedback for query reformulation (e.g., focus, expansion).

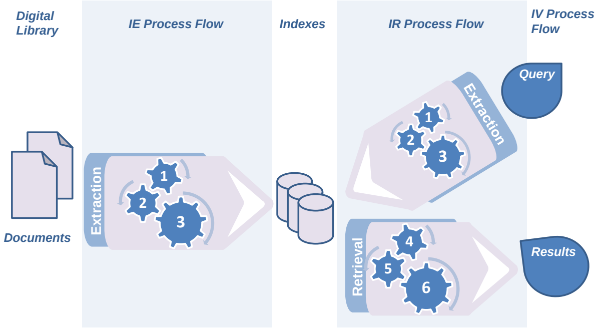


Fig. 1 Information extraction, retrieval, and visualization processes

GIR systems extract geographic information from documents: they aim to extract spatial (SF) and calendar / temporal (CF) features (i.e., entities) from texts, and then interpret them. For instance, 'River Thames' is tagged as an absolute SF (ASF) whereas 'North of the River Thames' is tagged as a relative SF (RSF) — spatial orientation relation [8]. In the same way, 'Spring 1840' is tagged as an absolute CF (ACF) whereas 'circa Spring 1840' is tagged as a relative CF (RCF) — temporal adjacency relation [9]. In the next section, we present common characteristics of IE, indexing, and retrieval processes supporting GIR systems.

2.1.1 Information extraction and indexing

IE processes extract either the entire information from a document (full-text indexing), or only specific parts (restrained to information matching predefined rules). In the former full-text case, extracted terms are weighted thanks to statistical approaches (all terms of a document are processed [11, chap. 2-4]). On the contrary, in the latter rule-based case, IE relies on predefined rules in order to extract specific information [12], which is not weighted. Once extracted, this information can be used for a variety of purposes, including text retrieval. Figure 1 illustrates specific information extraction processes supported by NLP. We may refer to Gate [13], LinguaStream [14], Miracle [15], SxPipe [16], and UIMA [17] NLP platforms; comparative results are available in [16]. LingPipe,[1] MetaCarta,[2] OpenCalais,[3] and OpenNLP[4] taggers support spatial named entity extraction from textual documents. In the GIR context, a process flow for textual content is usually comprised of the three following steps:

- *Named Entity Recognition* (Fig. 1, process 1): tokenization splits the document into smaller chunks of text. Then, these chunks are transformed into lexemes by means of lexical and morphological analysis. Finally named entities are recognized (NER).
- *Named Entity Validation* (Fig. 1, process 2): knowledge-based resources are used to validate candidate named entities (i.e., SFs and CFs).
- *Named Entity Interpretation* (Fig. 1, process 3): the syntactic analysis, based on grammar rules, allows relations between lexemes to be found. Finally, a semantic process consists in collecting meaningful lexeme groups. Knowledge-based resources are used to disambiguate and then associate representations to ASFs and ACFs. These resources are also used to analyze spatial and temporal relationships, and to compute corresponding RSF and RCF representations.

Validated named entities and their representations are stored into indexes for further retrieval.

2.1.2 Information retrieval

IR deals with models, techniques, and procedures to retrieve information that has already been processed, organized, and stored [11, chap. 5-6]. In a classical indexing approach, a weight is assigned to each term from a document (e.g., TF·IDF [18] weighting scheme). The same extraction and indexing process is applied to queries (Fig. 1, processes 1-3 in the retrieval stage). Usually, probabilistic [19] or vector-based models [20] are used to match documents with queries. For example, in a vector-based approach, the similarity between the query vector and the vector of each document in the collection is computed with equations involving such vectors (e.g., cosine). The corresponding relevance score is then used to rank the final list of retrieved documents in response to the query.

In the same way, GIR processes analyze user queries: spatial and / or temporal information is extracted and interpreted, so that user needs are represented by SFs and CFs. Then, the GIR system achieves the following tasks.

- *Parsing* (Fig. 1, process 4) spatial and temporal indexes in order to match the query representations with indexed documents.
- *Measuring similarity* (Fig. 1, process 5) between the query representations and the indexed documents (e.g., SFs and CFs).
- *Retrieving* (Fig. 1, process 6) a list of relevant documents.

Larson and Frontiera [21], as well as Andogah [22] review major techniques for measuring similarities. Andogah [22] defines spatial representations as 'spatial footprints' and describes three measures:

- *Euclidean distance* measures the proximity between document footprints and query footprints.
- *Extent of overlap* measures the proportion of overlap between document footprints and query footprints; the greater overlap, the higher the relevance of the document.
- *Containment relation* ranks documents by ratio of document footprints to query footprints (e.g., when the document footprints are inside the query footprints) or by ratio of query footprints to document footprints (e.g., when the query footprints are inside the document footprints).

Le Parc-Lacayrelle et al. [9], as well as Alonso et al. [23] describe similar measures proposed in temporal IR. Moreover, Kalczynski et al. [24] introduce *t-zoidal* fuzzy representations of specific temporal expressions in business news. They propose the 'temporal document retrieval model' and compute similarity measures (e.g., cosine) by relying on the Vector Space Model.

Although specific topical, spatial, and temporal IR techniques, models, and tools lead to improvements of traditional search engines, there are few works that combine these three

[1] http://alias-i.com/lingpipe
[2] http://www.metacarta.com
[3] http://www.opencalais.com
[4] http://opennlp.sourceforge.net

dimensions. The purpose of relevance ranking in GIR is to return a ranked list of documents satisfying topical, spatial, and temporal criteria altogether. Most of current works, however, only combine two of these dimensions. We may distinguish the three major following approaches:

- *Parallel filtering* computes intersections of result sets (i.e., topically, spatially, and temporally relevant documents).
- *Sequential filtering* processes spatial and / or temporal dimensions first, and then realizes term-based topical scoring on the remaining documents. Footprint-based filtering is the most often used combination approach (i.e., all spatially relevant documents are considered, then these documents are ranked according to their topical scores resulting from term-based IR) [21].
- *Linear interpolation* combines spatial and / or temporal scores with topical IR scores with a linear function, whereas weighted harmonic-mean supports weighted combination. Note that better performance is observed when the topical dimension outweighs the spatial dimension [22].

2.2 Literature review

Works related to GIR include the following prominent projects, presented by chronological order of publication:

1. GIPSY [25], 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 identify most representative geographic areas for indexing a document.
2. GEOSEM [26], for 'Geographic Semantic,' is dedicated processing the semantics of geographic information in documents (texts, maps, charts).
3. SPIRIT [27], for 'Spatially-Aware Information Retrieval on the Internet,' aims to find Web pages that refer to places or geographic areas specified in a query.
4. GRID [28], for 'Geospatial Retrieval of Indexed Documents,' is dedicated to textual IR. It combines keyword and areas of interest on a cartographic interface for search.
5. DIGMAP [29], for 'Discovering our Past World with Digitised Maps,' is dedicated to cultural and scientific heritage promotion, such as digital libraries of old maps.
6. GEOTRACKER [30], for 'Geospatial and Temporal RSS Tracking,' presents RSS feeds along with spatial and temporal (chronological) dimensions.
7. STEWARD [31], for 'Spatio-Textual Extraction on the Web Aiding Retrieval of Documents', performs extraction, retrieval, and geographic area visualization for unstructured texts.
8. PIV [8], for 'Virtual Itineraries in the Pyrenees Mountains,' is dedicated to cultural heritage promotion: newspapers, novels, and travelogues of the 19th century.
9. CITER [32], for 'Creation of a European History Textbook Repository,' offers history textbook retrieval.
10. GEOOREKA [33], for 'Geographically-enhanced web search engine,' focuses on textual documents (newspaper extracts).
11. SINAI [34], for 'Systemas intelligentes de acceso a la información,' is a GIR system managing textual documents (e.g., stories and newswire from newspapers).
12. Document Trajectory [35], for 'Document Trajectory Extraction,' explores Web textual documents (e.g., Wikipedia featured articles corpus) and presents event / location pairs in a chronological order.
13. Local Search [36], is a GIR system on textual documents, that uses an ontology-based index for the spatial dimension.

As shown in Tab. 2, these geographic information extraction, indexing, and retrieval systems handle the spatial criteria in priority (ASF, see Sect. 3.1).

Regarding the information extraction process, most widespread NLP platforms are Gate, LinguaStream, and UIMA. Only the Local Search GIR system integrates the LingPipe named entity recognition and classification system [37].

Regarding the information retrieval process, the CITER, Document Trajectory, GEOTRACKER, GRID, and SPIRIT systems use a 'parallel filtering' (PF) approach as the mode for criteria combination. They process simultaneously and separately each dimension, and then combine the different result lists by computing intersections. Because this PF operation is based on the Set theory, document relevance is boolean. As a result, the retrieved documents are not ranked. Other systems like SINAI and STEWARD, use a 'sequential filtering' (SF) approach. This consists in parsing the corpus for one dimension first, and then applying the other dimension filters on the document subset (e.g., priority to topical dimension, then to spatial dimension in STEWARD). Here the results are ranked according to their score for the last criterion (e.g., the spatial dimension for STEWARD). These filtering approaches can combine several criteria without any information standardization requirement. This is, however, not possible to compute a global score so as to rank the results according to all involved dimensions at the same time. Finally, systems like DIGMAP, GEOOREKA, GEOSEM, and Local Search use 'linear interpolation' (LI) approaches (e.g., arithmetic mean). Nevertheless, the global ranking score that they compute may favor lists with scores in higher ranges, as they do not standardize the combined criteria. Notice that, although relying on LI as well, our PIV system applies a standardization approach for spatial and temporal entities before computing combination algorithms.

Table 2 Comparison of GIR projects and systems with respect to NLP platforms used for information extraction, indexed geographical dimensions, and their combination modes during retrieval stage (LI: linear interpolation, PF: parallel filtering, SF: sequential filtering)

System	Year	Reference	IE	Index characteristics					IR
				Spatial		Temporal		Topical	
				ASF	RSF	ACF	RCF		
GIPSY	1994	[25]	–	+	−	−	−	−	−
GEOSEM	2003	[26]	LinguaStream	+	+	+	+	+	LI
SPIRIT	2005	[27]	Gate	+	−	−	−	+	PF
GRID	2006	[28]	Gate	+	−	−	−	+	PF
DIGMAP	2007	[29]	–	+	−	+	−	+	LI
GEOTRACKER	2007	[30]	Miracle	+	−	+	−	+	PF
STEWARD	2007	[31]	–	+	−	−	−	+	SF
PIV	2008	[8]	LinguaStream	+	+	+	+	+	LI
CITER	2009	[32]	Gate	+	−	+	−	+	PF
GEOOREKA	2009	[33]	–	+	−	−	−	+	LI
SINAI	2009	[34]	Gate	+	−	−	−	+	SF
Document Trajectory	2010	[35]	UIMA	+	−	+	−	−	PF
Local Search	2010	[36]	LingPipe	+	+	−	−	+	LI

We summed up in Tab. 2 the main characteristics of the indexes built by these systems. This study should be complemented by an evaluation of these systems according to their effectiveness in retrieving relevant documents, hence the need for the evaluation framework proposed in this article. In the next section, we introduce the PIV, as an appropriate case to study thanks to the evaluation framework we propose.

3 PIV GIR system

We designed and developed the PIV system in order to validate the aforementioned hypothesis: combining the three geographic dimensions improves retrieval effectiveness. This section describes characteristics of the PIV system.

The PIV system relies on a tile-based standardization of spatial and temporal information. It features a generic standardization resulting in an homogeneous representation of the different dimensions of geographic information: the tile. We apply discretization methods to segment the temporal axis or the spatial plan in tiles. This approach is in line with standard IR models based on term lemmatization / truncation and weighting schemes (e.g., term frequency in the Vector Space Model). So, we propose to generalize spatial and temporal information extracted from textual documents: the PIV system builds multi-scaled indexes (e.g., district tiling or city tiling for the spatial dimension). Then, three monodimensional IR systems (i.e., SpatialPIV denoted Sp, TemporalPIV denoted Te, and Terrier [38] denoted To, respectively spatial, temporal, and topical IR systems in next sections) produce ranked document lists. They produce comparable final result lists since representations and scores are standardized prior to a combination process.

This section introduces components of the PIV system: indexing process flows integrating standardization, and retrieval process involving combination of result lists.

3.1 Indexing: spatial, temporal, and topical process flows

3.1.1 Principles of multi-dimensional indexing

As proposed by Clough et al. [39], we process each of the three geographic dimensions independently. This can be achieved by building several indexes, one per dimension, as advised in [40]. In this way, one can restrain the search on one criterion while managing the indexes (e.g., allowing corpus evolution with document insertion and deletion). Our approach processes indexes independently and combines results later on for supporting multi-dimensional IR. It contributes to the GIR field as defined by Jones and Purves [41], as well as GIR in Digital Libraries as defined by Larson [42]. Figure 2 illustrates the three process flows dedicated to textual document indexing in the PIV system. As shown in Fig. 3, spatial and temporal indexes result from four main stages.

Lexico-syntactic process. This first stage consists in a lexico-syntactical processing sequence [8]: it addresses spatial feature (SFs, see Sect. 2) and calendar feature (CFs) extraction. This process is supported by the LinguaStream platform (i.e., LS in Tab. 2) [26, 43]. After a common textual tokenization preprocessing sequence, we adopt an 'active reading' behavior: a tagger of candidate tokens locates those corresponding to spatial or temporal information using typo-

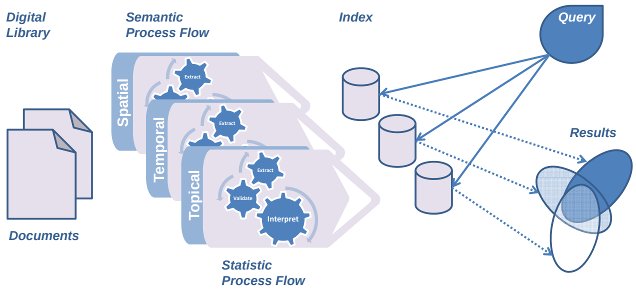


Fig. 2 PIV process flows dedicated to textual document corpora

graphic and lexical rules and lexicons. A morpho-syntactic analysis gathers tokens to constitute nominal groups corresponding to candidate SFs and CFs (e.g., *'River Thames,' 'Death Valley,' 'Palace of Versailles'*).

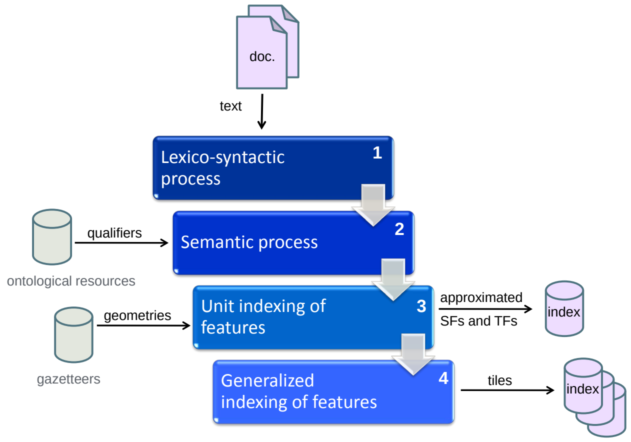


Fig. 3 The four main indexing stages of the PIV system

Semantic process. This second stage interprets SFs and CFs: it computes a symbolic representation of any SF and CF. We designed semantic patterns and implemented them with Definite Clause Grammars (DCGs): a recursive semantic analysis of each SF and CF matches absolute SFs (ASFs), relative SFs (RSFs), absolute CFs (ACFs) and relative CFs (RCFs). Therefore, anytime a spatial or temporal relationship is found out, the corresponding RSF and RCF is tagged (i.e., adjacency, inclusion, orientation, distance or union relationship with a SF or a CF). Similarly, a second set of DCGs is dedicated to the categorization of the SF and CF (e.g., oronym, hydronym, town, road tags characterize SFs; day, week, month, season, year, decade tags characterize CFs). For example, the nominal syntagm *'North of the River Thames'* is tagged as a RSF with an '*orientation spatial relationship:north*' tag and a '*category:hydronym*' tag. This RSF is recursively defined from the ASF labeled *River Thames*. Each of these tagging processes is supported by specific ontological resources [44].

Unit indexing of features. This third stage aims at interpreting symbolic representations: corresponding numeric approximations are computed. Specific algorithms compute time intervals for CFs [44] and geometries for SFs [8]. The spatial approximation process involves resources like the PostGIS[5] GIS, as well as gazetteers (e.g., French national geographic institute (IGN) BDNYME database®, GeoNames geographical database, local contributive resources) in order to validate and retrieve each ASF geometric representation. Then, spatial relationships (e.g., orientation:north) are processed and corresponding representations are approximated. Finally, GIS operators (e.g., translation, intersection) are applied to the original SF. Similarly, the temporal approximation process first associates a date or a time interval with each ACF. Then, temporal relationships (e.g., adjacency:around) are interpreted and one or several time intervals are computed from the original CF. A spatial 'raw' index and a temporal 'raw' index result from this third stage. The spatial index describes each SF with a corresponding geometry, nominal syntagm, paragraph, and document ID. Similarly, the temporal index describes each CF with a corresponding period, nominal syntagm, paragraph, and document ID. This first indexing level supports several IR scenarios: the overlapping area between spatial or temporal zones of the index and those of the query are computed, and then a set of ranked relevant items are returned [9, 45].

Generalized indexing of features. Finally, this fourth stage is dedicated to the standardization of unit indexes. First, a tiling process segments each geographic dimension into a corresponding spatial, temporal, and topical grid. Then, all the spatial, temporal, and topical entities described in the first level of index are grouped together, weighted, and mapped into the associated tiles (see Fig. 4 and Fig. 5). This approach consists in rearranging geographic information into a uniform representation: the tile. As a result, this approach requires complementary processes applied to these three dimensions:

1. *Spatial standardization* allows a grid-based or an administrative zoning (e.g., district, city, county) of the territory mentioned in the document collection and a projection of the SFs of the spatial raw index on this segmentation.
2. *Temporal standardization* allows a grid-based or a calendar cutting (e.g., day, month, year) of the period mentioned in the document collection and a projection of the CFs of the temporal raw index on this segmentation.
3. *Topical standardization* allows a segmentation (based on topics of domain specific ontologies) of the subjects mentioned in the document collection and a projection of the terms of an topical raw index on this segmentation.

The frequency of a tile corresponds to the number of SFs / CFs / terms that intersect it (i.e., a SF / CF / term can intersect several tiles). For example of Document$_1$ in Fig. 4

[5] http://postgis.refractions.net

illustrates SF ‘Paris,’ and RSF ‘South of Bobigny’ projection on two spatial segmentations: a grid-based representation and a city representation. Tiles T4, T5, T6, T8, and T9 are involved with the grid-based spatial zoning, while in the city spatial zoning only ‘Paris’ and ‘Créteil’ tiles are concerned.

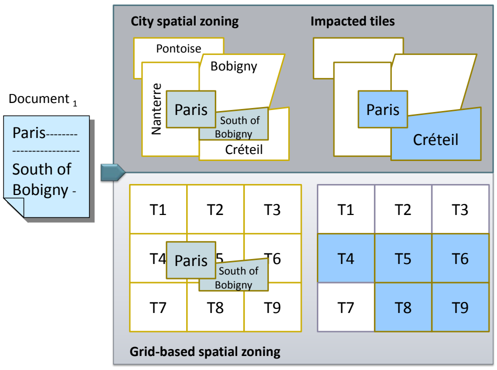


Fig. 4 Index standardization: a spatial tiling example

The Document$_1$ in Fig. 5 illustrates four CF projection on a temporal segmentation: a grid-based representation. Five tiles (T2, T3, T4, T5, and T7) are involved in such a grid-based temporal cutting.

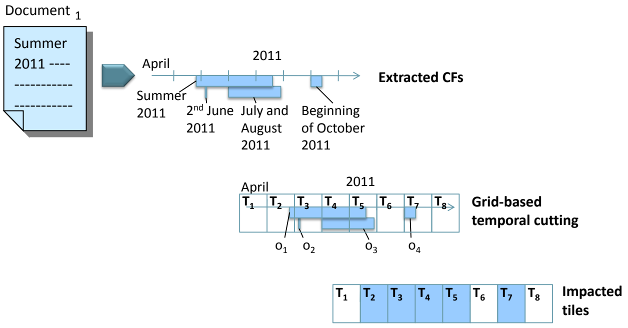


Fig. 5 Index standardization: a temporal tiling example

This standardization stage presents a twofold interest. It allows to express spatial, temporal, and topical representations into an homogeneous representation supported by spatial, temporal, and topical tile-based segmentations. Moreover, it allows the implementation of state-of-the-art models for computing IR relevance score based on such spatial, temporal, or topical tiles given their frequency in documents. Such a generalization (i.e., standardization) leading gathering geographic entities into tiles causes some loss of accuracy [46]. This is balanced, however, by the introduction of tile frequency computation, as well as the computation of indexes of different scales. For example, the city zoning index (county index) must be used by the IR system (IRS) if the query targets a city (a larger area). So, this fourth stage (in Fig. 3) produces several multi-scaled indexes for any spatial and temporal dimension. The interested reader may refer to [47] in this respect.

3.1.2 Per dimensional indexing performance

Experiments with the PIV system reported in [47] showed that the proposed spatial and temporal tile-based indexing approach is effective for GIR. Concerning the cultural heritage corpus used in [47], best performance was achieved by spatial tiling with city-based administrative zoning, and temporal tiling with month-based calendar cutting.

These indexing strategies may be associated with discrete scores or continuous scores depending on the ratio of overlap 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 (TFc for continuous Tile Frequency) yielded the best results in our context [48, 47]. For example, in Fig. 4, tile T5 has TF $= 2.00$ (TFc $= 0.67$ – sum of the ‘Paris’ SF and ‘South of Bobigny’ SF areas overlapping tile T5) whereas tile T6 has TF $= 1.00$ (TFc $= 0.28$). That is the reason why we use indexes corresponding to city-based administrative zoning and month-based calendar cutting, as well as TFc continuous frequency computing in Sect. 6 in order to evaluate different combination scenarios of PIV spatial and temporal retrieval results. However, as PIV topical process flow is not fully automated yet, we restrain PIV topical component to the state-of-the-art Terrier full-text IRS [38].

3.2 Retrieval: combination of result lists

As for any search engine, the PIV search engine requires the document collection to be indexed prior to be queried. As presented in the previous section, monodimensional indexes are built for each of the three dimensions. Then, PIV is able to process full-text geographic queries (e.g., *‘Potato famine in Ireland after the mid-19th century’*) as follows. First, the involved dimensions are identified in order to trigger off the corresponding indexes only. This leads to three result lists, or less when not every dimension is involved. Second, these independent results need to be combined together to return a comprehensive, single result list *l* to the user. In the IR literature, two approaches are promoted for combining several result lists into a single one. The combination may be realized according to document scores (Sect. 3.2.1) or document rankings in result lists (Sect. 3.2.2).

Table 3 Combination functions proposed by Fox and Shaw [49]

Combining function	Definition
CombMIN	minimum of individual similarities
CombMAX	maximum of individual similarities
CombSUM	summation of individual similarities
CombANZ	CombSUM ÷ number of nonzero similarities
CombMNZ	CombSUM × number of nonzero similarities

3.2.1 Combination according to document scores

Fox and Shaw [49] introduced to the IR field the Comb* combining functions shown in Tab. 3. The resulting combined list l gathers distinct documents retrieved by the source IRSs together. For instance, CombMNZ returns the similarity s of a document d in l 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 the score for d in the result lists of the source IRSs, the more relevant in l it shall be (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. Note that we showed the effectiveness of Comb* functions in previous works involving the combination of topical and semantic dimensions [50].

In addition, Lee [51] compared CombMNZ with other combining functions on TREC[6] test collections, and demonstrated its effectiveness. That is the reason why we experimented with this function for combining monodimensional IRS results. As each source IRS may compute scores in a different numeric domain, we normalized them within $[0,1]$ according to (equ. 1) as proposed by Lee [51].

$$\text{normalized\_similarity} = \frac{\text{unnormalized\_similarity} - \text{min\_similarity}}{\text{max\_similarity} - \text{min\_similarity}} \tag{1}$$

We illustrate a combination example using CombMNZ in Fig. 6 (a–c). We consider, for query $q = 8$, the results retrieved by the three monodimensional 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. Combinations of these IRS results is detailed in Fig. 6 (d–e). It shows CombMNZ similarity values and the corresponding computation details in Fig. 6 (d). These similarity values are based on the normalized values of IRS sources, see Fig. 6 (a–c). This basic example 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. In addition, 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.

3.2.2 Combination according to document ranks

Result combination has also been supported by 'voting' algorithms designed as election methods [52, 53, 54]. They have been applied to IR for metasearch engines [55, 56], which merge results of several search engines according to document ranks. The prominent algorithm for this purpose is Borda Count [52], and works as follows according to Aslam and Montague [55]:

> "Each voter ranks a fixed set of c candidates in order of preference. For each voter, the top ranked candidate is given c points, the second ranked candidate is given $c-1$ points, and so on. If there are some candidates left unranked by the voter, the remaining points are divided evenly among the unranked candidates. The candidates are ranked in order of total points, and the candidate with the most points wins the election." [55]

In the IR context, the *candidates* are all the documents from the corpus (e.g., $c = 1{,}000{,}000$), and *voters* are source IRSs. For any query, there are obviously more unranked documents than ranked documents. According to the aforementioned Borda algorithm, these unranked documents would get the same divided amount of points leading to large extent of tied documents. This leads to a twofold issue. First, from a user perspective, documents classified as irrelevant (i.e., unranked) by the IRS are wrongly attributed credit. Second, from a system perspective, these tied documents introduce a bias in IR evaluation [57].

The result of Borda Count for the previous example is shown in Fig. 6 (e). Note that we kept Borda Count principles, and made the following changes due to aforementioned issues. First, the top ranked candidate is given n points, where n is the length of the longest result list. Second, documents unranked get no points. The interested reader may refer to [58, 59] for a more detailed coverage of result combination based on document ranks.

4 State of the art of GIR evaluation

We introduce in Sect. 4.1 key concepts related to IR evaluation. Then we focus in Sect. 4.2 on existing IR and GIR evaluation frameworks.

4.1 Concepts: campaign, task, test collection

IR has a long tradition of evaluation [60, 61, 62]. Organizing a *campaign* is a way to evaluate IRSs; it is comprised of six main steps:

[6] Text REtrieval Conference, see `http://trec.nist.gov`.

(a) Topical IRS

q	d	s
8	d_4	14.5
8	d_3	12
8	d_5	8.7
8	d_1	0.5

(b) Spatial IRS

q	d	s
8	d_6	150
8	d_1	120
8	d_4	80
8	d_7	-10
8	d_2	-30

(c) Temporal IRS

q	d	s
8	d_6	1
8	d_4	0.7
8	d_7	0.5
8	d_1	0.5
8	d_2	0.5

Combination of result lists

(d) IRS combination with normalized CombMNZ

q	d	Combined similarity s
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_6	$4.0000 = 2\cdot\left(0 + \frac{150+30}{150+30} + \frac{1-0.5}{1-0.5}\right)$
8	d_1	$2.5000 = 3\cdot\left(\frac{0.5-0.5}{14.5-0.5} + \frac{120+30}{150+30} + \frac{0.5-0.5}{1-0.5}\right)$
8	d_3	$0.8214 = 1\cdot\left(\frac{12-0.5}{14.5-0.5} + 0 + 0\right)$
8	d_5	$0.5857 = 1\cdot\left(\frac{8.7-0.5}{14.5-0.5} + 0 + 0\right)$
8	d_7	$0.2222 = 2\cdot\left(0 + \frac{-10+30}{150+30} + \frac{0.5-0.5}{1-0.5}\right)$
8	d_2	$0.0000 = 2\cdot\left(0 + \frac{-30+30}{150+30} + \frac{0.5-0.5}{1-0.5}\right)$

(e) IRS combination with Borda count

q	d	Combined similarity s
8	d_4	$12 = 5+3+4$
8	d_6	$10 = 0+5+5$
8	d_1	$8 = 2+4+2$
8	d_7	$5 = 0+2+3$
8	d_3	$4 = 4+0+0$
8	d_5	$3 = 3+0+0$
8	d_2	$2 = 0+1+1$

Fig. 6 Illustration of result lists combination with (d) normalized CombMNZ [49] versus (e) Borda count [52]

1. Organizers spread a call for participation, which presents the proposed IR tasks. For instance, an *ad hoc* task requires to retrieve a list of relevant documents for a given query. In contrast, a *question answering* task requires the retrieval of a piece of information for a given query. For the query *'States in the USA'* we would get a list of documents dealing with this subject for *ad hoc* task, whereas we would get the list of the states in the USA for *question answering* task.
2. IRS designers register to the tasks their are interested in. These are then referred to as *participants*.
3. Organizers provide a *corpus* of documents, and 25+ *topics* representing information needs (i.e., detailed queries with description and narrative).
4. Participants process the corpus, submit the topics to their IRS, and then send obtained results, also known as *runs* (e.g., per topic document list ranked by decreasing relevance) to the organizers.
5. Organizers constitute the *relevance judgments*: the set of relevant documents for each topic. Then, organizers check participants' run against these relevance judgments by means of predefined appropriate measures. The computed value represents the effectiveness (i.e., measurement of result quality) of the IRS for the considered topic. Aggregating all the scores obtained by the IRS for each of the 25+ topics (e.g., averaging over them) leads to a global evaluation score for the IRS.
6. Organizers publish participants' results and generally release the *test collection* (i.e., corpus, topics, and relevance judgments). This collection may then be reused for evaluating IRSs later.

In the IR evaluation literature, especially for TREC [63] campaigns, 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 natural 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 [64], note that 50 topics is standard at TREC.
2. The *corpus* of documents, some of which are relevant for the proposed topics. A regular TREC corpus for classical *ad hoc* task is made up of 800,000 documents and more [63].
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, a human assessor skims through each document for evaluating whether it matches the information need corresponding to topic t or not: the document is then marked as *relevant* or *non-relevant*.

4. The *measures* allowing evaluation of any IRS, which retrieves a list of pairs (d,s) representing the relevance score s of each retrieved document d for topic t. 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 [11, chap. 8].

4.2 Existing evaluation frameworks

As stated by Voorhees [65] and Sanderson [62], IR evaluation is classically based on the Cranfield evaluation paradigm [60]. This evaluates IR system effectiveness according the topical dimension of document and queries. Several prominent evaluation frameworks implemented this paradigm, such as TREC [66, 67], NTCIR [68, 69], CLEF [**?** 70], and INEX [71, 72]. These offer the evaluation of various task types in many languages (e.g., English, Japanese, Spanish).

Further initiatives considered other dimensions than the topical one (Tab. 4). The TEMPEVAL [73] evaluation framework is concerned with the temporal dimension. Building on both of these initiatives, Bucher et al. [74] proposed to evaluate two dimensions at the same time: spatial and topical dimensions. This proposal was realized in the GeoCLEF [75] task of the CLEF program [**?**]. It notably allowed Perea-Ortega et al. [76] to evaluate the effectiveness of classical topical IRSs, such as Lemur [77], Lucene [78], and Terrier [38]. Finally, GeoTime [79] considers spatial and temporal dimensions, whereas GikiCLEF [80] considers topical and spatial dimensions.

Table 4 Dimensions evaluated by GIR systems evaluation frameworks

Year	Framework	Measured dimensions		
		Topical	Spatial	Temporal
1992	TREC [66]	✓		
1999	NTCIR [68]	✓		
2000	CLEF [**?**]	✓		
2002	INEX [71]	✓		
2005	Bucher et al. [74]	✓	✓	
2005	GeoCLEF [75]	✓	✓	
2009	GeoTime [79]		✓	✓
2009	GikiCLEF [80]	✓	✓	
2009	TEMPEVAL [73]			✓
2010	Palacio et al. [81]	✓	✓	✓

To the best of our knowledge, GIR contributions (reviewed in Sect. 2) 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.

5 Proposed 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 5.1 details the design of a test collection covering the three geographic dimensions; then Sect. 5.2 reports the analysis of the PIV GIR system, enabling us to assess its effectiveness.

5.1 GIR test collection

Several test collections were used by 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 design a test collection 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 Ireland after mid-19th century'* and its narrative may be *'Relevant documents mention scarcity of food and its consequences in Ireland after 1849.'*
2. A *corpus* covering the three dimensions: documents convey not only the usual topical dimension but also additional spatial and temporal dimensions.
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 birthplace. Although spatially relevant, it does not match the query *'Pubs in Dublin.'* Such a subtlety requires the assessment of the global match between the query and the document.
   In judging a given document, the assessor evaluates its adequacy according to each of the three dimensions. 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' [74].
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 (from 1 to 3), and global relevance (4). No assumption was made regarding the relative importance of dimensions; they were equitably considered.

4. *The NDCG measure (Normalized Discounted Cumulative Gain)* [82] suited for gradual qrels associated with the three dimensions of geographic information. Notice that *AP* and *MAP* are not appropriate because they are limited to binary qrels. *NDCG* was notably used at TREC-9 for the Web task [63]. It 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.
5. *Geographic resources* georeferencing spatial entities that occur in the corpus.

The experimental protocol for measuring the effectiveness of GIR systems is detailed in the next section. GIR systems are evaluated on the basis of the *runs* they provided (i.e., the retrieved document list per topic).

5.2 Protocol for GIR systems 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 relevance. Indeed, the evaluation framework allows effectiveness evaluation for the following IRSs:

- monodimensional (i.e., topical To, spatial Sp, and temporal Te),
- bidimensional (i.e., To+Sp, To+Te, and Sp+Te allowing the measurement of effectiveness improvement according to each missing dimension),
- and GIR systems combining the three dimensions (i.e., To+Sp+Te).

Building on TREC experience, 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 of significance on the paired observed differences are also reported with respect to significance *p*-values resulting from Student's paired (the difference is computed between paired values m_i^t and m_j^t) two-tailed (because $\forall t \in [1,n]\ m_i^t \not\gtrless m_j^t$) t-test. Although this test theoretically requires a normal data distribution, Hull [83] 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 said to be statistically significant; the smaller the *p*-value (e.g. $p < 0.01$, $p < 0.001$), the more significant the difference is [83].

6 Case study involving the evaluation framework

As a case study, we applied the evaluation framework presented in Sect. 5 for evaluating the PIV system. We consider in this section the constituted test collection, the experiments conducted with various combining functions, the comparative analyses that we carried out (global and per topic levels), and their limitations.

6.1 Design of the MIDR_2010 test collection

The MIDR_2010 test collection (Online Resource 1) is comprised of the four components identified in Sect. 5.1.

1. The *corpus* represents 5,645 paragraphs extracted from 11 books published between the 18th and 20th centuries, and belonging to the Aquitaine Regional Library. They were scanned and processed with OCR software. As retrieved by the IRS, a document *d* is one of these paragraphs; it is considered as the best entry point in its associated book.
2. We collected 41 *topics* covering the three dimensions of geographic information.
3. The *qrels* were obtained by querying three IRSs — a topical IRS based on PL2 IR model (built-in Terrier [38] configuration), a spatial IRS, and a temporal IRS — with the 'title' part of the topics.
4. The *geographic resources* corresponding to the corpus are provided by the French National Geographic Institute (BD NYME® database) and by a contributive local gazetteer.

For each topic, the results retrieved by the IRSs were considered for setting up the pool (see Sect. 4.1). It was then assessed according to binary judgments for each dimension plus global judgment. These four judgments were aggregated in a single gradual value, as presented in Sect. 5.1.

6.2 Performance assessment of PIV

As explained in Sect. 3.2, PIV combines the results of three monodimensional IRSs with either Comb* or Borda combining functions. Among Comb* functions, Lee [51] reported that normalized CombMNZ yields best performance on TREC3 dataset. In this section, we check whether this observation holds for the MIDR_2010 collection by comparing the performances of the five functions considered in [51], whose definitions are shown in Tab. 3. We also check the performance of these functions against the performance of Borda Count. As measured with *NDCG*, we report in Tab. 5 the performance of combining functions applied to the three dimensions (topical, spatial, and temporal). These are computed according to unnormalized vs. normalized input scores in order to assess the effect of normalization ($< 6\%$). Our experiment corroborates Lee's observations [51] since the optimal function over Comb* is normalized CombMNZ (0.7977). Notice that it is slightly outperformed by Borda Count (0.8043). The limited 0.8% improvement is, however, not statistically significant ($p = 0.347$), which means that one function may perform better than the other only by chance. In addition, further experiments showed that CombMNZ results are better than Borda Count when using a weighted variant, as discussed in Sect. 6.2.3.

Table 5 Combining function performance (*NDCG*) for unnormalized and normalized input scores

Combining function	Input		Gain (%)
	Unnormalized	Normalized	
CombMIN	0.6045	0.6372	5.41
CombMAX	0.7512	0.7580	0.91
CombSUM	0.7776	0.7830	0.69
CombANZ	0.7385	0.7355	−0.41
CombMNZ	0.7858	0.7977	1.51
Borda Count	**0.8043**	–	–

In the next section, we compare the performance of various monodimensional IRS combinations with normalized CombMNZ.

6.2.1 Comparative analysis of IRS effectiveness

In Tab. 6, we report the observed comparisons between various IRSs and two baselines identified in [76]: To^+ is a strong baseline corresponding to OkapiBM25 model; To^- is a weaker baseline corresponding to TF·IDF model. In addition, Sp denotes the spatial IRS, and Te denotes the temporal IRS (see Sect. 5.2). The reported results were computed with latest trec_eval[7] (version 9.0) used at TREC. They show effectiveness of search engines according to the 41 tested topics.

Overall, the two baselines showed identical effectiveness, contrary to the observations presented in [76]. This 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 yield better performance.

Table 6 Effectiveness of monodimensional and combined IRSs with respect to topical baseline To^+. The * symbol denotes a significant difference ($p < 0.001$) compared with baseline To^+

# IRSs	Monodimensional IRS				*MANDCG*	Gain To^+
	To^-	To^+	Sp	Te		(%)
①	✓				0.4796	0.0
		✓			0.4796	0.0
			✓		0.4922	2.6
				✓	0.4841	0.9
②	✓	✓			0.4797	0.0
	✓		✓		0.6407*	33.6
	✓			✓	0.7063*	47.3
		✓	✓		0.6411*	33.7
		✓		✓	0.7068*	47.4
			✓	✓	0.7182*	49.7
③	✓	✓	✓		0.6357*	32.5
	✓	✓		✓	0.6890*	43.7
	✓		✓	✓	0.7964*	66.1
		✓	✓	✓	**0.7977***	**66.3**
④	✓	✓	✓	✓	0.7694*	60.4

Regarding monodimensional IRSs ① they all achieve similar performance; best effectiveness (0.4922) is reached by the spatial IRS. In addition, combining at least two heterogeneous dimensions ② yields better performance. Notice that the associated improvement is statistically significant regarding the baseline, Te+Sp being the most effective combinations (0.7182). However, the combination To^++Te is similar in effectiveness (0.7068). 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').

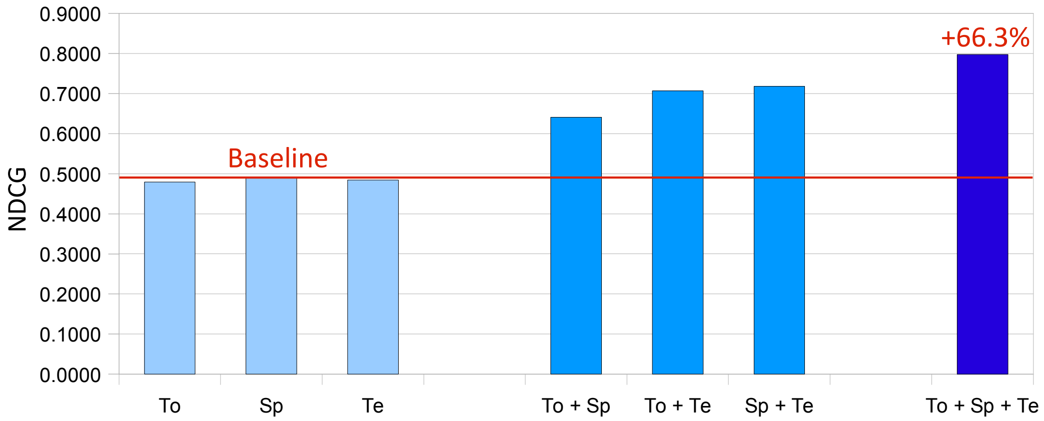
Combining the three dimensions ③ (0.7977) yields better results (+11.1%) than the best bidimensional combination (0.7182). The difference is statistically significant ($p = 0.000$). Adding to these dimensions a topical IRS ④ (To^+ or To^-) does not result in more improvement (0.7694). The resulting topical reinforcement may lessen the complementary information contributed by the two other dimensions.

To sum up, as shown in Fig. 7, combining the three dimensions provides the best performance (0.7977). The 66.3% improvement regarding To^+ validates the hypothesis formulated in this article: the combination of the three geographic information dimensions yields better performance than considering only the topical dimension.

6.2.2 Per topic analysis of the PIV system

Having identified the combination of IRSs yielding best performance for PIV (To^++Sp+Te), we analyze in this section

[7] http://trec.nist.gov/trec_eval



Topical	✓				✓	✓			✓
Spatial		✓			✓		✓		✓
Temporal			✓			✓	✓		✓

Fig. 7 Comparative *NDCG* results for each geographic information dimension, and associated best combinations with normalized CombMNZ

its per topic performance. This thorough analysis aims at refining the aforementioned observations. For that purpose, Tab. 7 reports for each topic:

1. The topic number (column 1),
2. The number of documents judged as relevant (see *qrels*) for each of the three dimensions (columns 2–4),
3. The PIV performance (*NDCG*) for each topic (column 5),
4. The number of relevant documents retrieved by only one, by only two, or by the three IRSs, each IRS being associated with one dimension (columns 6–8).

We observed that the repartition of relevant documents varies over the three dimensions for each topic. Nevertheless, the spatial dimension is predominant. This is due to a specificity of the MIDR_2010 collection, as it covers topics related to the Pyrenean area.

Moreover, the low standard deviation $\sigma_{NDCG} = 0.0616$ of *NDCG* values shows a stable per topic performance. Overall, the performance variations with respect to the average $NDCG = 0.7977$ computed against the 41 topics is in the $[-25\%, +12\%]$ range.

Finally, we studied the proportion of relevant documents contributed by each dimension. Were the dimensions redundant in providing relevant documents (all retrieving the same documents) or complementary to each other (each one retrieving specific documents)? We found a few (10 on average) relevant documents retrieved by the three dimensions at the same time (intersection), as shown in the last column of Tab. 7. This proves the following point: the three dimensions are well complementary to each other.

6.2.3 Effects of dimension weighting on effectiveness

In previous sections, the three dimensions were combined equitably with CombMNZ (i.e., each dimension has the same influence as the two other ones). We wondered whether a weighted combination could lead to improve the effectiveness of the system. As a result, we investigate the effects of dimension weighting on the GIR system effectiveness in this section. Let us introduce three parameters defined on $W = \{0.0, \ldots, 1.0\}$. These are $(\alpha, \beta, \gamma) \in W^3$, such that a weight of 1.0 is shared among parameters, that is $\alpha + \beta + \gamma = 1.0$. We performed a weighted combination according to (equ. 2), which results in the L result list.

$$\begin{aligned} L \leftarrow \mathit{CombMNZ}\big(\alpha \cdot \mathit{normalize}(\mathsf{To}^{+}), \\ \beta \cdot \mathit{normalize}(\mathsf{Sp}), \\ \gamma \cdot \mathit{normalize}(\mathsf{Te})\big) \end{aligned} \tag{2}$$

Overall, 66 weighting configurations comply with the aforementioned constraints. We report the five best and five worst performing configurations in Tab. 8, as well as the associated gain over the topical baseline To^{+}. Monodimensional IRSs are still the worst performing systems. Combination of two or more dimensions, whatever the weights, always yield better effectiveness. The best weighted configuration (0.8341) overcomes the unweighted CombMNZ configuration (0.7977) by a 4.5% increase. The temporal dimension ($\gamma = 0.8$) is the most influential dimension among the two others $(0.1 + 0.1)$.

7 Discussion

Limited combination policy. As explained in the previous section, CombMNZ can be extended so as to give a specific weight to each dimension. Setting a value for these

Table 7 Per topic study of the relevant document distribution according to the three dimensions, performance of PIV (*NDCG*), and complementarity of the three dimensions

Topic	# relevant documents			*NDCG*	Complementarity		
	To$^+$	Sp	Te		1 dim	2 dim	3 dim
1	67	260	134	0.7715	148	134	15
2	60	213	138	0.8756	195	87	14
3	42	111	101	0.7808	201	25	1
4	16	16	74	0.5960	82	9	2
5	93	85	34	0.8169	144	34	2
6	48	60	59	0.8496	139	14	2
7	71	24	93	0.7354	127	29	1
8	27	177	103	0.8098	229	39	1
9	41	209	112	0.7949	154	95	6
10	68	110	102	0.7703	176	52	6
11	33	111	6	0.7137	90	30	6
12	41	246	116	0.7638	150	101	17
13	106	120	47	0.7683	201	33	2
14	74	110	107	0.7857	137	74	2
15	89	152	103	0.7010	190	74	2
16	74	40	0	0.8072	84	15	2
17	197	116	23	0.8315	190	70	2
18	37	141	112	0.8758	218	36	2
19	113	175	101	0.7163	199	83	8
20	180	11	73	0.7327	147	57	1
21	13	150	122	0.7083	157	58	4
22	7	162	121	0.8882	198	46	4
23	126	114	120	0.8360	205	58	13
24	170	70	30	0.8160	185	41	1
25	234	102	110	0.8195	132	100	38
26	160	227	131	0.7918	71	126	65
27	19	41	113	0.8580	149	12	65
28	21	94	117	0.8326	139	45	1
29	47	41	120	0.7749	149	25	3
30	51	91	131	0.8751	213	30	3
31	16	194	55	0.8496	185	40	3
32	48	136	96	0.8955	218	31	3
33	145	25	120	0.7965	185	42	7
34	15	179	126	0.8229	158	72	6
35	108	181	55	0.8015	161	84	5
36	5	224	123	0.7510	160	93	2
37	23	60	9	0.7552	84	4	2
38	84	104	80	0.8406	190	39	2
39	60	122	118	0.8781	210	42	2
40	48	140	131	0.8555	178	66	3
41	170	253	120	0.7627	95	116	72
Avg	74	127	92	0.7977	162	55	10

parameters requires, however, a training phase. It would be interesting to study how to dynamically adjust weights during search sessions. Moreover, for a given dimension, it might be interesting to be able to set a ratio for a first criterion while setting another one for a second criterion (e.g., 0.3 for temporal criterion #1, and 0.1 for temporal criterion #2).

Restriction to French language. A limitation of the PIV system is related to its dependency to French language. The rules supporting the tagging of spatial and temporal relationships need to be adapted and extended for dealing with multilingual corpora.

Table 8 Five best and worst performing weighting configurations. The * symbol denotes a significant difference ($p < 0.001$) compared with baseline To$^+$

Weights			*NDCG*	Improvement over To$^+$ (%)
To$^+$ (α)	Sp (β)	Te (γ)		
0.1	**0.1**	**0.8**	**0.8341**	**73.9***
0.1	0.2	0.7	0.8309	73.2*
0.1	0.3	0.6	0.8253	72.1*
0.2	0.1	0.7	0.8249	72.0*
0.2	0.2	0.6	0.8222	71.4*
0.1	0.9	0.0	0.6327	31.9*
0.9	0.1	0.0	0.6321	31.8*
0.0	1.0	0.0	0.4922	2.6
0.0	0.0	1.0	0.4841	0.9
1.0	0.0	0.0	0.4796	0.0

Entity ambiguity. Disambiguation processes driven by the semantic analysis of results might be improved for managing worldwide spatial areas. As the geographic zone concerned by our corpus is limited to the Pyrenees Mountains, and the corresponding period covers the 19th century, we faced few disambiguation problems. In order to analyze different and larger areas and periods, the semantic analysis of spatial and temporal entities should be supported by qualifiers (e.g., hydronym, oronym, road, city or day, month, season) computed by current PIV modules; this kind of information may guide disambiguation processes effectively.

Limited corpus size. The experiment reported in this article is limited regarding the number of documents in the MIDR_2010 test collection (5,645 paragraphs for a 3.7 MB total size). A larger corpus needs to be constituted for conducting further experiments.

8 Conclusion and future work

In this article, we considered geographic IRSs dedicated to spatial, temporal, and topical information. As common search engines show limitations in such contexts, we put forward the hypothesis that processing these three dimensions improves the accuracy of retrieval results. In order to verify this hypothesis, we defined an evaluation framework for evaluating GIR systems according to the three dimensions of geographic information. As a case study, we applied it with the MIDR_2010 test collection (Online Resource 1). We used this test collection for validating different combining approaches with the PIV GIR system [8]. We showed a 73.9% improvement over a state-of-the-art topical baseline, this performance gain being statistically significant. A thorough per topic study of the results showed that the three dimensions are not redundant, but they complement each other.

These results provide an empirical validation of our proposals experimented with the PIV GIR system. In addition, this evaluation framework supports:

1. Comparison of different indexing processes (e.g., tiling variants, weighting schemes) applied to each of these three dimensions (e.g., experiments of various types of tiling approaches for the spatial dimension). Integrating other dimensions is also possible, such as confidence in information and information freshness [84].
2. Comparison of alternative combining operators. Thus, it is possible to confront the performance of combination operators based on the relevance scores of the documents (e.g., Comb* [49]) with that of operators based on document rankings (e.g., Borda Count [52]).
3. Comparison of our PIV GIR system with state-of-the-art GIR systems identified in Sect. 2.
4. Evaluation of the influence of dimension weighting with PIV. This would enable us to adjust our model so as to better simulate human perception, based on the priorities established mentally between dimensions (e.g., preference to topical dimension). Likewise, it would enable us to evaluate the constraint-based combination approach associating requirements and preferences to any search criterion, which we plan to investigate in our future works.

We are currently investigating constraint-based search (allowing to express requirements and preferences) according to linear approaches [e.g., 85], harmonic approaches [e.g., 86], fuzzy approaches with OWA operators [e.g., 87, 88], or other proposals like Choquet's integral [e.g., 89]. Furthermore, we plan to promote the proposed evaluation framework in GIR tasks of evaluation campaigns. It would allow the measurement of system effectiveness according to the full range of geographic information involving topical, spatial, and temporal dimensions.

Acknowledgements We are indebted to Valérie Bougeant, Jean-Paul Cabanac, Christian Chevalier, Benoît Chirle, Firas Damak, Agnès Itier, Rudy Kapler, Anaïs Lefeuvre, Nicolas Noullet, and Camille Raymond for their invaluable and voluntary contributions regarding document assessments. The PIV project was funded by the Pau City Council and the MIDR multimedia library.

Appendix

Table 9 Abbreviations used in this article

Abbreviation	Meaning
ACF	Absolute Calendar Feature
AP	Average Precision
ASF	Absolute Spatial Feature
CF	Calendar Feature
DCG	Definite Clause Grammar
GIR	Geographic Information Retrieval
GIS	Geographic Information System
IE	Information Extraction
IR	Information Retrieval
IV	Information Visualization
MAP	Mean Average Precision
NDCG	Normalized Discounted Cumulative Gain
NER	Named Entity Recognition
NLP	Natural Language Processing
RCF	Relative Calendar Feature
RSF	Relative Spatial Feature
SF	Spatial Feature
TF	Term Frequency
TFc	continuous Term Frequency

References

1. Jon Purday. Think culture: Europeana.eu from concept to construction. *The Electronic Library*, 27(6): 919–937, 2009. ISSN 0264-0473. doi: 10.1108/02640470911004039.
2. E. Lynn Usery. A feature-based geographic information system model. *Photogramm. Eng. Rem. Sens.*, 62(7): 833–838, 1996. ISSN 0099-1112.
3. Mark Sanderson and Janet Kohler. Analyzing Geographic Queries. In *SIGIR-GIR'04: Proceedings of the Workshop on Geographic Information Retrieval at SIGIR*, 2004.
4. Saeid Asadi, Chung-Yi Chang, Xiaofang Zhou, and Joachim Diederich. Searching the World Wide Web for Local Services and Facilities: A Review on the Patterns of Location-Based Queries. In Wenfei Fan, Zhaohui Wu, and Jun Yang, editors, *WAIM'05: Proceedings of the 6th International Conference on Web-Age Information Management*, volume 3739 of *LNCS*, pages 91–101. Springer, 2005. ISBN 3-540-29227-6. doi: 10.1007/11563952_9.
5. Ligiane A. Souza, Jr. Davis, Clodoveu A., Karla A. V. Borges, Tiago M. Delboni, and Alberto H. F. Laender. The Role of Gazetteers in Geographic Knowledge Discovery on the Web. In *LA-WEB'05: Proceedings of the 3rd Latin American Web Congress*, pages 157–165. IEEE Computer Society, October 2005. doi: 10.1109/LAWEB.2005.38.

6. Qingqing Gan, Josh Attenberg, Alexander Markowetz, and Torsten Suel. Analysis of geographic queries in a search engine log. In *LocWeb'08: Proceedings of the first international workshop on Location and the web*, pages 49–56, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-160-6. doi: 10.1145/1367798.1367806.
7. Rosie Jones, Wei V. Zhang, Benjamin Rey, Pradhuman Jhala, and Eugene Stipp. Geographic intention and modification in web search. *Int. J. Geogr. Inf. Sci.*, 22(3):229–246, 2008. ISSN 1365-8816. doi: 10.1080/13658810701626186.
8. Mauro Gaio, Christian Sallaberry, Patrick Etcheverry, Christophe Marquesuzaa, and Julien Lesbegueries. A global process to access documents' contents from a geographical point of view. *J. Vis. Lang. Comput.*, 19(1): 3–23, 2008. ISSN 1045-926X. doi: 10.1016/j.jvlc.2007.08.010.
9. Annig Le Parc-Lacayrelle, Mauro Gaio, and Christian Sallaberry. La composante temps dans l'information géographique textuelle. *Document Numérique*, 10(2): 129–148, 2007. ISSN 1279-5127. doi: 10.3166/dn.10.2.129-148.
10. Christian Sallaberry, Mustapha Baziz, Julien Lesbegueries, and Mauro Gaio. Towards an IE and IR System Dealing with Spatial Information in Digital Libraries – Evaluation Case Study. In Jorge Cardoso, José Cordeiro, and Joaquim Filipe, editors, *ICEIS'07: Proceedings of the 9th International Conference on Enterprise Information Systems*, pages 190–197, 2007. ISBN 978-972-8865-92-4.
11. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, July 2008. ISBN 978-0521865715.
12. Robert J. Gaizauskas. Recent Advances in Computational Terminology edited by Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme. *Comput. Linguist.*, 29(2):328–332, 2003. ISSN 0891-2017. doi: 10.1162/coli.2003.29.2.328.
13. Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: an architecture for development of robust HLT applications. In *ACL'02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 168–175, Morristown, NJ, USA, 2002. ACL. doi: 10.3115/1073083.1073112.
14. Frédérik Bilhaut and Antoine Widlöcher. Linguastream: an integrated environment for computational linguistics experimentation. In *EACL'06: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 95–98, Morristown, NJ, USA, 2006. ACL. doi: 10.1.1.134.4336.
15. Zhu Liu, David C. Gibbon, and Behzad Shahraray. Multimedia Content Acquisition and Processing in the MIRACLE System. In *CCNC'06: Proceedings of the 3rd IEEE Conference on Consumer Communications and Networking*, pages 272–276, January 2006. ISBN 1-4244-0085-6. doi: 10.1109/CCNC.2006.1593030.
16. Benoit Sagot and Pierre Boullier. SxPipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues*, 49(2):155–188, 2008.
17. David Ferruci and Adam Lally. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10 (3-4):327–348, 2004. ISSN 1351-3249. doi: 10.1017/S1351324904003523.
18. Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *J. Doc.*, 28(1): 11–21, 1972. ISSN 0022-0418. doi: 10.1108/eb026526.
19. Stephen E. Robertson and Karen Spärck Jones. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.*, 27 (3):129–146, 1976. doi: 10.1002/asi.4630270302.
20. Gerard Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Commun. ACM*, 18(11): 613–620, November 1975. doi: 10.1145/361219.361220.
21. Ray R. Larson and Patricia Frontiera. Spatial Ranking Methods for Geographic Information Retrieval (GIR) in Digital Libraries. In *ECDL'04: Proceedings of the 8th European Conference on Digital Libraries*, volume 3232 of *LNCS*, pages 45–56. Springer, 2004. ISBN 3-540-23013-0. doi: 10.1007/978-3-540-30230-8_5.
22. Geoffrey Andogah. *Geographically Constrained Information Retrieval*. PhD thesis, University of Groningen, Netherlands, May 2010.
23. Omar Rogelio Alonso. *Temporal information retrieval*. PhD thesis, University of California, Davis, May 2008.
24. Pawel Jan Kalczynski and Amy Chou. Temporal document retrieval model for business news archives. *Inf. Process. Manage.*, 41(3):635–650, 2005. ISSN 0306-4573. doi: 10.1016/j.ipm.2004.01.002.
25. Allison Gyle Woodruff and Christian Plaunt. Gipsy: automated geographic indexing of text documents. *J. Am. Soc. Inf. Sci.*, 45(9):645–655, 1994. ISSN 0002-8231. doi: 10.1002/(SICI)1097-4571(199410)45:9<645::AID-ASI2>3.0.CO;2-8.
26. Frédérik Bilhaut, Thierry Charnois, Patrice Enjalbert, and Yann Mathet. Geographic reference analysis for geographic document querying. In *HLT-NAACL'03: Proceedings of the workshop on Analysis of geographic references*, pages 55–62, Morristown, NJ, USA, 2003. ACL. doi: 10.3115/1119394.1119403.
27. Subodh Vaid, Christopher B. Jones, Hideo Joho, and Mark Sanderson. Spatio-textual Indexing for Geographical Search on the Web. In *SSTD'05: Proceed-*

ings of the 9th international Symposium on Spatial and Temporal Databases, volume 3633 of *LNCS*, pages 218–235. Springer, 2005. ISBN 3-540-28127-4. doi: 10.1007/11535331_13.

28. Valcartier. GRID – geospatial retrieval of indexed document. Technical report, R&D pour la défense, Canada, 2006.
29. Bruno Martins, Jose Borbinha, Gilberto Pedrosa, João Gil, and Nuno Freire. Geographically-aware information retrieval for collections of digitized historical maps. In *GIR'07: Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 39–42, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-828-2. doi: 10.1145/1316948.1316959.
30. Yih-Farn Robin Chen, Giuseppe Di Fabbrizio, David Gibbon, Serban Jora, Bernard Renger, and Bin Wei. Geotracker: geospatial and temporal RSS navigation. In *WWW'07: Proceedings of the 16th international conference on World Wide Web*, pages 41–50, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242579.
31. Michael D. Lieberman, Hanan Samet, Jagan Sankaranarayanan, and Jon Sperling. STEWARD: Architecture of a Spatio-Textual Search Engine. In *GIS'07: Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, pages 1–8, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-914-2. doi: 10.1145/1341012.1341045.
32. Dieter Pfoser, Alexandros Efentakis, Thanasis Hadzilacos, Sophia Karagiorgou, and Giorgos Vasiliou. Providing Universal Access to History Textbooks: A Modified GIS Case. In *W2GIS'09: Proceedings of the 9th International Symposium on Web and Wireless Geographical Information Systems*, volume 5886 of *LNCS*, pages 87–102, 2009. ISBN 978-3-642-10600-2. doi: 10.1007/978-3-642-10601-9_7.
33. Davide Buscaldi and Paolo Rosso. Geooreka: Enhancing Web Searches with Geographical Information. In Valeria De Antonellis, Silvana Castano, Barbara Catania, and Giovanna Guerrini, editors, *SEBD'09: Proceedings of the Seventeenth Italian Symposium on Advanced Database Systems*, pages 205–212. Edizioni Seneca, 2009. ISBN 978-88-6122-154-3.
34. Miguel Á. García-Cumbreras, José M. Perea-Ortega, Manuel García-Vega, and L. Alfonso Ureña-López. Information retrieval with geographical references. Relevant documents filtering vs. query expansion. *Inf. Process. Manage.*, 45(5):605–614, 2009. ISSN 0306-4573. doi: 10.1016/j.ipm.2009.04.006.
35. Jannik Strötgen, Michael Gertz, and Pavel Popov. Extraction and exploration of spatio-temporal information in documents. In *GIR'10: Proceedings of the 6th Workshop on Geographic Information Retrieval*, pages 16:1–16:8, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-826-1. doi: 10.1145/1722080.1722101.
36. Nieves Brisaboa, Miguel Luaces, Ángeles Places, and Diego Seco. Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based index. *GeoInformatica*, 14 (3):307–331, 2010. ISSN 1384-6175. doi: 10.1007/s10707-010-0106-3.
37. Nicola Stokes, Yi Li, Alistair Moffat, and Jiawen Rong. An empirical study of the effects of NLP components on Geographic IR performance. *Int. J. Geogr. Inf. Sci.*, 22(3):247–264, January 2008. ISSN 1365-8816. doi: 10.1080/13658810701626210.
38. Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. Terrier: A high performance and scalable information retrieval platform. In *OSIR'06: Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval*, 2006.
39. Paul Clough, Hideo Joho, and Ross Purves. Judging the Spatial Relevance of Documents for GIR. In *ECIR'06: Proceedings of the 28th European Conference on IR Research*, volume 3936 of *LNCS*, pages 548–552. Springer, 2006. ISBN 3-540-33347-9. doi: 10.1007/11735106_62.
40. Bruno Martins, Mário J. Silva, and Leonardo Andrade. Indexing and ranking in Geo-IR systems. In *GIR'05: Proceedings of the 2005 workshop on Geographic information retrieval*, pages 31–34, New York, NY, USA, 2005. ACM. ISBN 1-59593-165-1. doi: 10.1145/1096985.1096993.
41. Christopher B. Jones and Ross Purves. GIR'05 2005 ACM workshop on geographical information retrieval. *SIGIR Forum*, 40(1):34–37, 2006. ISSN 0163-5840. doi: 10.1145/1147197.1147202.
42. Ray R. Larson. Geographic Information Retrieval and Digital Libraries. In *ECDL'09: Proceedings of the 13th European Conference on Digital Libraries*, volume 5714 of *LNCS*, pages 461–464. Springer, 2009. doi: 10.1007/978-3-642-04346-8_59.
43. Antoine Widlöcher and Frédérik Bilhaut. La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus. In *TALN'05: Actes de la 12^e Conférence sur le Traitement Automatique du Langage Naturel*, 2005.
44. Eric Kergosien, Mouna Kamel, Christian Sallaberry, Marie-Noëlle Bessagnet, Nathalie Aussenac-Gilles, and Mauro Gaio. Construction automatique d'ontologie et enrichissement à partir de ressources externes. In *JFO'09: Actes des 3^e Journées Francophones sur les Ontologies*, pages 11–20, 2009. ISBN 978-1-60558-842-1.
45. Christian Sallaberry, Mauro Gaio, Damien Palacio, and Julien Lesbegueries. Fuzzying GIS Topological Functions for GIR Needs. In *GIR'08: Proceeding of the 2nd international workshop on Geographic informa-*

tion retrieval, pages 1–8, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-253-5. doi: 10.1145/1460007.1460008.

46. Nicholas R. Chrisman. Deficiencies of sheets and tiles: building sheetless databases. *Int. J. Geogr. Inf. Sci.*, 4(2): 157–167, 1990. doi: 10.1080/02693799008941537.
47. Damien Palacio, Christian Sallaberry, and Mauro Gaio. Normalizing Spatial Information to Improve Geographical Information Indexing and Retrieval in Digital Libraries. In *ISGIS'10: Proceedings of the Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science proceedings*, pages 229–234, 2010.
48. Damien Palacio, Christian Sallaberry, and Mauro Gaio. Normalizing Spatial Information to Better Combine Criteria in Geographical Information Retrieval. In *ECIR-GIIW'09: Proceeding of the international workshop on Geographic Information on the Internet*, pages 37–48, 2009.
49. Edward A. Fox and Joseph A. Shaw. Combination of Multiple Searches. In Donna K. Harman, editor, *TREC-1: Proceedings of the First Text REtrieval Conference*, pages 243–252, Gaithersburg, MD, USA, February 1993. NIST.
50. Gilles Hubert and Josiane Mothe. An adaptable search engine for multimodal information retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 60(8):1625–1634, 2009. ISSN 1532-2882. doi: 10.1002/asi.21091.
51. Joon Ho Lee. Analyses of Multiple Evidence Combination. In *SIGIR'97: Proceedings of the 20th annual international ACM SIGIR conference*, pages 267–276, New York, NY, USA, 1997. ACM Press. ISBN 0-89791-836-3. doi: 10.1145/258525.258587.
52. Jean-Charles de Borda. Mémoire sur les élections au scrutin. In *Histoire de l'Académie Royale des Sciences*, Paris, 1781.
53. Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Imprimerie Royale, Paris, 1785.
54. Donald Saari. Which is better: the Condorcet or Borda winner? *Soc. Choice Welfare*, 26(1):107–129, 2006. ISSN 0176-1714. doi: 10.1007/s00355-005-0046-2.
55. Javed A. Aslam and Mark Montague. Models for metasearch. In *SIGIR'01: Proceedings of the 24th annual international ACM SIGIR Conference*, pages 276–284, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6. doi: 10.1145/383952.384007.
56. Mark Montague and Javed A. Aslam. Condorcet fusion for improved retrieval. In *CIKM'02: Proceedings of the 11th international conference on Information and knowledge management*, pages 538–548, New York, NY, USA, 2002. ACM. ISBN 1-58113-492-4. doi: 10.1145/584792.584881.
57. Guillaume Cabanac, Gilles Hubert, Mohand Boughanem, and Claude Chrisment. Tie-breaking Bias: Effect of an Uncontrolled Parameter on Information Retrieval Evaluation. In Maristella Agosti, Nicola Ferro, Carol Peters, Maarten de Rijke, and Alan F. Smeaton, editors, *CLEF'10 : Proceedings of the 1st Conference on Multilingual and Multimodal Information Access Evaluation*, volume 6360 of *LNCS*, pages 112–123. Springer-Verlag, September 2010. doi: 10.1007/978-3-642-15998-5_13.
58. Mohamed Farah and Daniel Vanderpooten. An Outranking Approach for Rank Aggregation in Information Retrieval. In *SIGIR'07: Proceedings of the 30th annual international ACM SIGIR conference*, pages 591–598, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277843.
59. Bing Liu. Information Retrieval and Web Search. In *Web Data Mining*, Data-Centric Systems and Applications, chapter 6, pages 183–236. Springer, 2007. ISBN 978-3-540-37882-2. doi: 10.1007/978-3-540-37882-2_6.
60. Cyril W. Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. ASLIB Cranfield Research Project, ASLIB, Cranfield, UK, October 1962.
61. Stephen Robertson. On the history of evaluation in IR. *J. Inf. Sci.*, 34(4):439–456, 2008. ISSN 0165-5515. doi: 10.1177/0165551507086989.
62. Mark Sanderson. Test collection based evaluation of information retrieval systems. *Found. Trends Inf. Retr.*, 4(4):247–375, 2010. ISSN 1554-0669. doi: 10.1561/1500000009.
63. Donna K. Harman. The TREC Test Collections. In Voorhees and Harman [67], chapter 2, pages 21–53. ISBN 0-262-22073-3.
64. Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *SIGIR'00: Proceedings of the 23rd international ACM SIGIR conference*, pages 33–40, New York, NY, USA, 2000. ACM. doi: 10.1145/345508.345543.
65. Ellen M. Voorhees. The philosophy of information retrieval evaluation. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *CLEF'01: Proceedings of the Second Workshop of the Cross-Language Evaluation Forum*, volume 2406 of *LNCS*, pages 355–370. Springer, 2002. ISBN 3-540-44042-9. doi: 10.1007/3-540-45691-0_34.
66. Donna K. Harman, editor. *TREC-1: Proceedings of the First Text REtrieval Conference*, Gaithersburg, MD, USA, February 1993. NIST.
67. Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval.* MIT Press, Cambridge, MA, USA, 2005. ISBN 0-262-22073-3.

68. Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Soichiro Hidaka. Overview of IR Tasks at the First NTCIR Workshop. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 11–44. NACSIS, 1999. ISBN 4-924600-77-6.
69. Noriko Kando. Evaluation of information access technologies at the NTCIR workshop. In Carol Peters, Julio Gonzalo, Martin Braschler, and Michael Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems*, volume 3237 of *LNCS*, pages 197–221. Springer, 2004. doi: 10.1007/978-3-540-30222-3_4.
70. Carol Peters and Martin Braschler. European Research Letter – Cross-Language System Evaluation: the CLEF Campaigns. *J. Am. Soc. Inf. Sci. Technol.*, 52(12):1067–1072, 2001. ISSN 1532-2882. doi: 10.1002/asi.1164.
71. Norbert Fuhr, Norbert Gövert, Gabriella Kazai, and Mounia Lalmas, editors. *INEX'02: Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)*, 2002.
72. Gabriella Kazai, Mounia Lalmas, Norbert Fuhr, and Norbert Gövert. A report on the first year of the initiative for the evaluation of XML retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 55(6):551–556, 2004. doi: 10.1002/asi.10386.
73. Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. The TempEval challenge: identifying temporal relations in text. *Lang. Resour. Eval.*, 43(2): 161–179, 2009. ISSN 1574-020X. doi: 10.1007/s10579-009-9086-z.
74. Bénédicte Bucher, Paul Clough, Hideo Joho, Ross Purves, and Awase Khirni Syed. Geographic IR Systems: Requirements and Evaluation. In *ICC'05: Proceedings of the 22nd International Cartographic Conference*. Global Congressos, 2005. ISBN 0-958-46093-0. CDROM.
75. Fredric C. Gey, Ray R. Larson, Mark Sanderson, Hideo Joho, Paul Clough, and Vivien Petras. GeoCLEF'05: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In *CLEF'05: Proceedings of the 5th workshop on Cross-Language Evalution Forum*, volume 4022 of *LNCS*, pages 908–919. Springer, 2006. ISBN 3-540-45697-X. doi: 10.1007/11878773_101.
76. José M. Perea-Ortega, Miguel A. García-Cumbreras, Manuel García-Vega, and L. A. Ureña-López. Comparing Several Textual Information Retrieval Systems for the Geographical Information Retrieval Task. In *NLDB'08: Proceedings of the 13th international conference on Natural Language and Information Systems*, pages 142–147, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-69857-9. doi: 10.1007/978-3-540-69858-6_15.
77. Paul Ogilvie and James P. Callan. Experiments Using the Lemur Toolkit. In *TREC'01: Proceedings of the 9th Text REtrieval Conference*, Gaithersburg, MD, USA, February 2001. NIST.
78. Otis Gospodnetić and Erik Hatcher. *Lucene in Action*. Manning Publications, 2005. ISBN 978-1-932-39428-3.
79. Fredric Gey, Ray Larson, Noriko Kando, Jorge Machado, and Tetsuya Sakai. NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search. In *NTCIR'10: Proceedings of the 8th NTCIR Workshop*, pages 147–153, Tockyo, Japan, 2010. NII. ISBN 978-4-86049-053-9.
80. Diana Santos and Luís Cabral. GikiCLEF: Expectations and Lessons Learned. In Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Djamel Mostefa, Anselmo Penas, and Giovanna Roda, editors, *CLEF'09: Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum*, volume 6241 of *LNCS*, pages 212–222. Springer Berlin / Heidelberg, 2010. doi: 10.1007/978-3-642-15754-7_23.
81. Damien Palacio, Guillaume Cabanac, Christian Sallaberry, and Gilles Hubert. Measuring Effectiveness of Geographic IR Systems in Digital Libraries: Evaluation Framework and Case Study. In *ECDL'10: Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*, volume 6273 of *LNCS*, pages 340–351. Springer, September 2010. doi: 10.1007/978-3-642-15464-5_34.
82. Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002. ISSN 1046-8188. doi: 10.1145/582415.582418.
83. David Hull. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *SIGIR'93: Proceedings of the 16th annual international ACM SIGIR conference*, pages 329–338, New York, NY, USA, 1993. ACM Press. ISBN 0-89791-605-0. doi: 10.1145/160688.160758.
84. Célia da Costa Pereira, Mauro Dragoni, and Gabriella Pasi. Multidimensional relevance: A new aggregation criterion. In *ECIR'09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 264–275, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-00957-0. doi: 10.1007/978-3-642-00958-7_25.
85. Mohamed Farah and Daniel Vanderpooten. An outranking approach for information retrieval. *Inf. Retr.*, 11 (4):315–334, 2008. ISSN 1386-4564. doi: 10.1007/s10791-008-9046-z.
86. Geoffrey Andogah and Gosse Bouma. Relevance measures using geographic scopes and types. In *CLEF'07: Proceedings of the 7th Workshop of the Cross-Language Evaluation Forum*, volume 5152 of *LNCS*, pages 794–801, 2008. ISBN 978-3-540-85759-4. doi: 10.1007/978-3-540-85760-0_100.

87. Ronald R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans. Syst. Man Cybern.*, 18(1):183–190, 1988. ISSN 0018-9472. doi: 10.1109/21.87068.
88. Mohand Boughanem, Yannick Loiseau, and Henri Prade. Refining aggregation functions for improving document ranking in information retrieval. In *SUM'07: Proceedings of the 1st international conference on Scalable Uncertainty Management*, pages 255–267, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-75407-7. doi: 10.1007/978-3-540-75410-7_19.
89. Christophe Labreuche and Michel Grabisch. The Choquet integral for the aggregation of interval scales in multicriteria decision making. *Fuzzy Sets Syst.*, 137 (1):11–26, 2003. ISSN 0165-0114. doi: 10.1016/S0165-0114(02)00429-3.