Text Meets Space: Geographic Content Extraction, Resolution and Information Retrieval

Jochen L. Leidner, Bruno Martins,
Katherine McDonough and Ross S. Purves

Tutorial held at ECIR 2020

Lisbon, Portugal, 14 April 2020
(online)

Tutorial Outline

- I. Geography and text
- II. Toponym recognition and resolution
- **III. Geographic relevance and ranking**
- IV. Applications
- V. Future challenges

Part III. Outline (1)

- **Introduction and motivation**
  - Classic retrieval systems
  - Extensions for spatio-textual search
  - GIR interfaces
    - Query formulation
    - Search result analysis
- **Spatio-textual queries**
  - Spatial queries
    - Proximity
    - Topology
    - Orientation/direction
  - Vagueness, ambiguity and the need for relevance ranking

Part III. Outline (2)

- **Relevance ranking for spatio-textual search**
  - Overview on text indexing and ranking
    - TF-IDF, BM25 and inverted indexes
  - Learning to rank from multiple signals
  - Relevance signals for GIR
  - GIR query processing
    - Overview on spatial data indexes
    - Combining spatial indexes with inverted lists
    - Combining thematic and spatial similarity
  - Diversity and other ranking topics
- **Putting it all together**
  - Exemplar systems
  - Evaluation campaigns
    - GeoCLEF, GeoTime and GikiCLEF
  - Future challenges

Introduction and Motivation

Introduction and Motivation

- Geographical information is pervasive.
  - Associating things and events to places.

- Users's information needs are often best approached from a geographical perspective.
  - Early studies estimated that 20% of Web searches were *local* in nature.
  - Much higher estimates for systems including map-based interfaces.
  - Particularly important in specialized domains.
    - Historians requiring data on specific areas (at particular times).
    - Domain scientists requiring regional/spatial data.

- Geographic Information Retrieval (GIR) is concerned with providing access to geo-referenced information sources.
  - Repositories of resources with explicit/exact geo-spatial references.
  - Repositories of **documents with implicit location references**.

Classic Retrieval Systems and Local Search

Local search functionalities are already commonly deployed together with standard text retrieval systems.

- Local search on SEs like Google (*neural matching towards DB entries*);
- Search over structured knowledge bases containing spatial information;
- Digital library interfaces for materials with geographic content/associations;
- Map-based interfaces for news search and media monitoring;
- Most of these mechanisms are in fact closer to **data retrieval**.

Despite previous research, some functionalities are still not commonly seen on commercial search products:

- Explicitly handling spatial queries;
- Geographic information retrieval over textual resources;
- Search results ranking combining thematic/spatial relevance.

Spatio-Textual Search and GIR

- Non-geographic subject restricted to a place
  - *Music festivals in Germany*
- Geographic subject with non-geographic restrictions
  - *Rivers with vineyards*
- Geographic subject restricted to a place
  - *Cities in Germany*
- Non-geographic subject associated to a place
  - *Independence of Quebec*
- Non-geographic subject that is a complex function of place
  - *European football cup matches*
- Vague and/or imprecise geographic region
  - *Climate in Sub-Saharan Africa*
- Geographical relations among places
  - *Oil and gas extraction between the UK and the Continent*
- Relations between events which require their precise localization
  - *Casualties in fights in Nagorno-Karabakh*

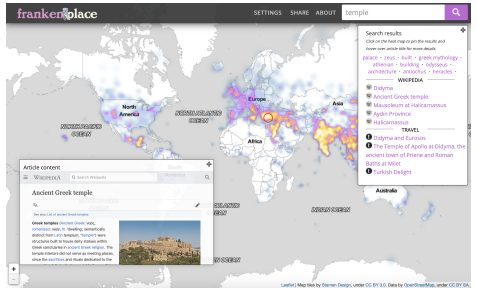
- Non-geographic subject restricted to a place
  - *Music festivals in Germany*
- Geographic subject with non-geographic restrictions
- G
- N
- N
- V
- Geographical relations among places
  - *Oil and gas extraction between the UK and the Continent*
- Relations between events which require their precise localization
  - *Casualties in fights in Nagorno-Karabakh*

Challenging examples for current technology!

- Standard text retrieval.
- Natural language question answering (*through neural models*).
- Reasoning over knowledge bases (*geo-spatial reasoning*).
- **Geographic information retrieval:**
  - Toponym resolution and document geo-coding;
  - Indexes combining textual and spatial information;
  - Geo-spatial query processing;
  - Geo-spatial relevance ranking;
  - Interfaces for interacting with results.

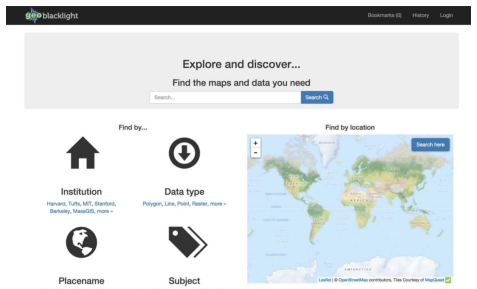
Query formulation:

- Input forms or specialized query syntax;
- Parsing natural language queries;
- Map-based interfaces for query formulation;
- Dynamic suggestions.



Search results analysis:

- Summarize results on dynamic maps;
  - E.g., keywords over maps;
- Present ranked lists of results:
  - Filter through spatial constraints;
  - Rank according to proximity/overlap;
  - Combine thematic/spatial similarity;
- Cluster results by location;
- Spatial diversification of ranked results.



Spatio-Textual Queries

Spatial and Spatio-Textual Queries

- Spatial queries involve relationships (e.g., containment or proximity) between regions located in space.
- Assume space is delineated by some coordinate system.
  - Distance and direction measured on a continuous scale.
  - Can use geometric and topological relationships.

- Three primary classes of (geo-)spatial requests from users:
  - Resources for a given location: **what's here?**
  - Locations related to a given resource: **where's this?**
  - Resources matching both particular themes and locations.

theme *related-to* **location**

- Within this classification, different query types can be distinguished by how the relations and locations are defined.

Proximity Queries

- Retrieve data within some distance of a given query point/region (*reference location* or *example object*).
  - Find resources on a given theme that relate to locations that are within a specified distance of a given city boundary.
- Rather than specifying an exact distance, users may specify a relation of *near* with no precise constraint.
  - Context dependent and challenging for automated interpretation.
- Processing proximity queries can involve reduction to topological queries (i.e., *containment within buffer regions*).

Topological Queries

- Retrieve data through restrictions on the nature of the connectivity between a pair of locations (e.g., geometries).
  - Find resources whose geographic scope lies *inside* the boundary of a given region (e.g., a specified country).

- Common relations:
  - Inside (conversely also contains);
  - Meets or touches;
  - Overlaps;
  - Equals;
  - Disjoint.

Orientation/Direction Queries

- Retrieve data through restrictions on directional/angular relationship between a pair of locations (e.g., geometries).
  - Find resources on some theme located to the *north of* a given region (e.g., a specified city).
- Processing these queries often also involves reduction to topological operations (i.e., *containment within regions*).
- Directional relations are vague and context dependent.
  - There is no agreed standard interpretation.
  - Less used in practical implementations.

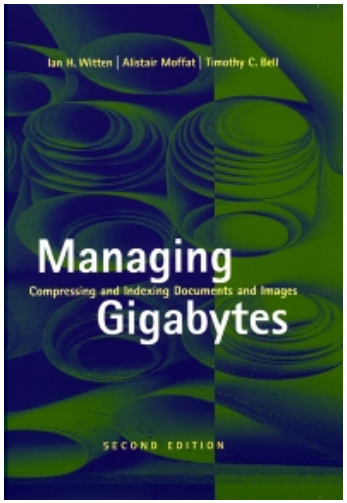
Vagueness, Ambiguity and Relevance Ranking

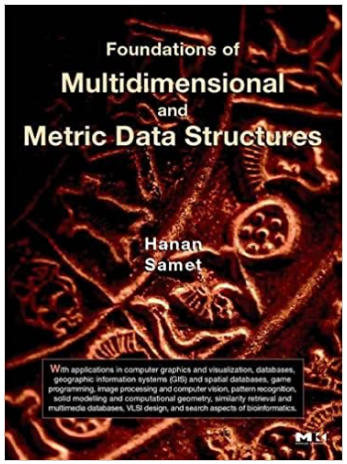
- Queries can involve vague and/or imprecise regions.
  - Vernacular toponyms.
  - Events often used as regions.
- Spatial relations can be expressed quantitatively.
  - Expressing distance with terms such as *near*.
  - Expressing direction with terms such as *north* and *in front*.
- Use combinations of the aforementioned restrictions.
  - Qualitative geo-spatial reasoning.
- Uncertainty in query interpretation.
  - Feature explicit theme *related-to* location.
  - Parsing spatial roles in natural language.

Relevance Ranking for Spatio-Textual Search

Relevance Ranking for Spatio-Textual Queries

- Queries can involve **thematic** and **spatial** restrictions.
- Query processing involves the use of **indexes**:
  - Efficiently retrieve candidate resources matching (hard or soft/relaxed) query restrictions:
    - Boolean operators over query terms;
    - Spatial constraints (e.g., topological);
  - Get **information needed for ranking**.
- Many challenges related to efficiency:
  - Index construction for very large collections;
  - Compressing the index (e.g., posting lists);
  - Query processing by manipulating posting lists;
  - Efficient top-k retrieval through pruning.





An Example Document Collection

Doc.	Text
1	That government is best which governs least
2	That government is best which governs not at all
3	The mass of men serve the state not as men, but as machines
4	Wooden men can be manufactured to serve the purpose
5	Government is at best but an expedient
6	But most governments are usually inexpedient

An Inverted Index (With Term Frequency Information)

Lexicon

Num.	Term	Doc. Freq.
1	best	3
2	expedient	1
3	government	4
4	governs	2
5	inexpedient	1
6	least	1
7	machines	1
8	manufactured	1
9	mass	1
10	men	2
11	purpose	1
12	serve	2
13	state	1
14	wooden	1

Inverted File

Inverted list
$(1;1),(2;1),(5;1)$
$(5;1)$
$(1;1),(2;1),(5;1),(6;1)$
$(1;1),(2;1)$
$(6;1)$
$(1;1)$
$(3;1)$
$(4;1)$
$(3;1)$
$(3;2),(4;1)$
$(4;1)$
$(3;1),(4;1)$
$(3;1)$
$(4;1)$

Ranked Retrieval for Text

- Boolean restrictions addressed through set operations.
- Docs. can be conceptually represented as term vectors:
  - $\vec{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$
  - $w_{i,j}$ is the weight of term i in document j
- Queries are also conceptually represented as vectors:
  - $\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$
- Vector operations can be used to compare queries with documents (e.g., dot product between normalized vectors).

How can we define the term weights?

1. Frequent terms within a document are more important.
2. Rare terms within a collection are more discriminative.
3. Long documents are more likely to contain different terms.

Defining Term Weights — TF and IDF

Let N be the total number of documents in the collection and n_i be the number of documents in which term k_i appears.

Importance within a document

The **normalized term frequency** of a term k_i in document d_j is given by:

$$\text{tf}_{i,j} = \frac{\text{freq}_{i,j}}{\max_l \text{freq}_{l,j}}$$

In the previous equation, $\text{freq}_{i,j}$ is the number of occurrences of term k_i in document d_j from the set of N documents.

Importance within a document collection

The **inverse document frequency** of a term k_i is given by:

$$\text{idf}_i = \log\left(\frac{N}{n_i}\right)$$

TF-IDF

The weight of a term k_i in document d_j according to the original vector space model can be given by the **tf-idf** formula:

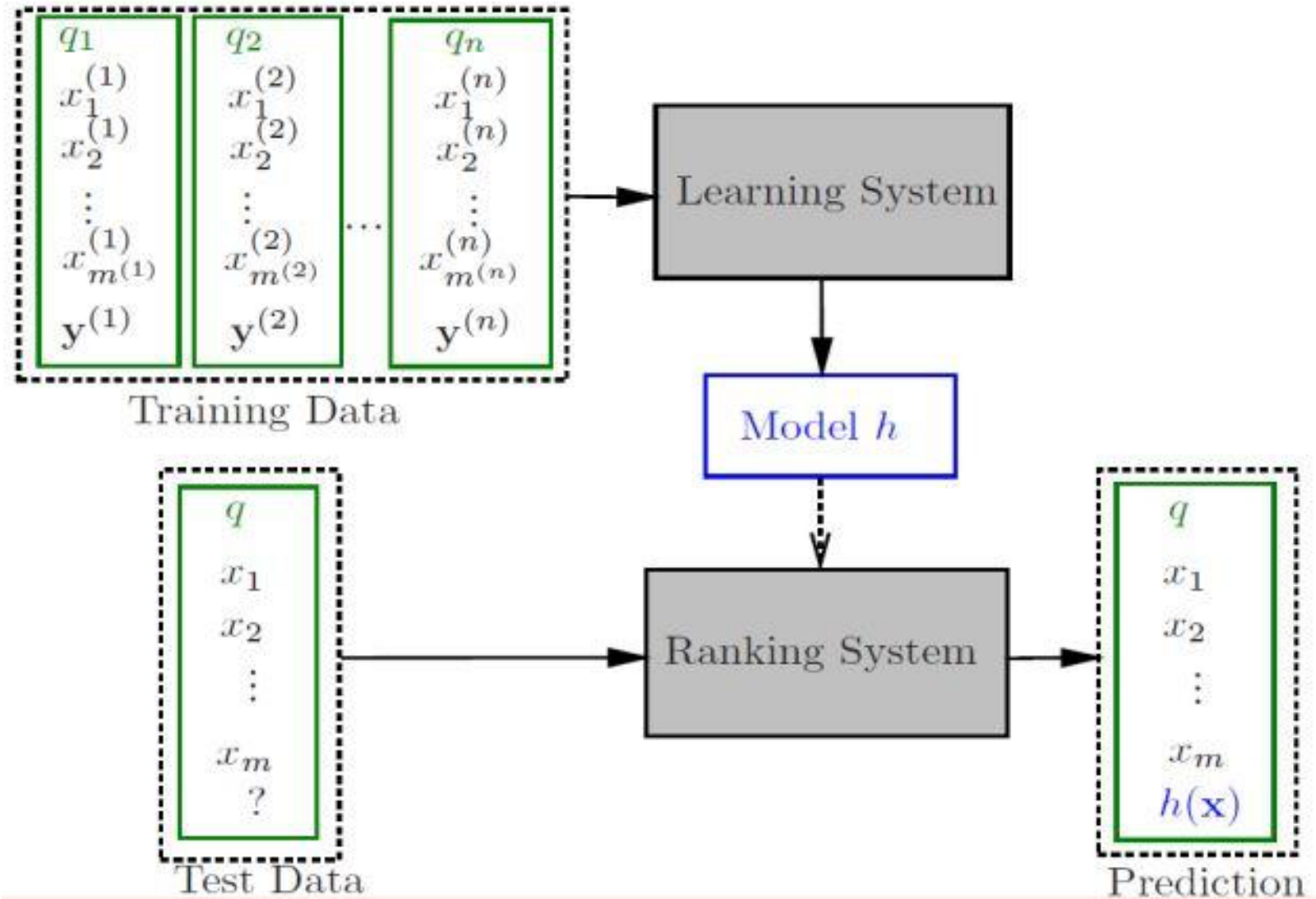
$$w_{i,j} = \frac{\text{freq}_{i,j}}{\max_l \text{freq}_{l,j}} \times \log\left(\frac{N}{n_i}\right)$$

BM25

We can consider not only the term frequency and inverse document frequency, but also the **document length as a normalization factor** for the term frequency.

$$\text{sim}(d_j, q) = \sum_{i \in q} \frac{\text{freq}_{i,j} \times (k_1 + 1)}{\text{freq}_{i,j} + k_1 \times \left(1 - b + b\frac{|d_j|}{\text{avgdl}}\right)} \times \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)$$

Learning to Rank

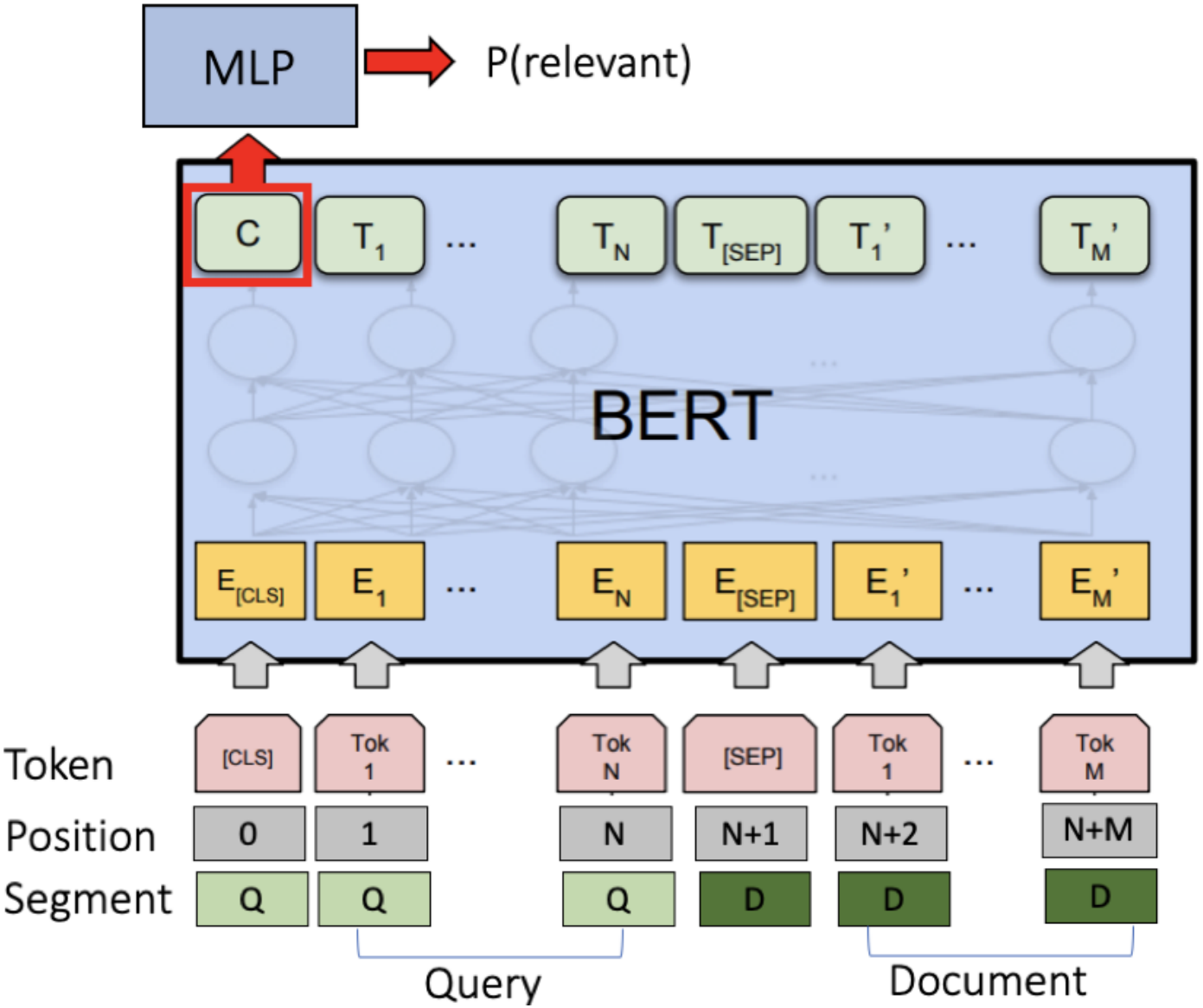


Ranking Signals for GIR

- Signals capturing thematic relevance and prior importance:
  - TF, IDF, TF-IDF, BM25, language modeling scores, ...
  - PageRank, clicks measured over query logs, ...

- **Signals capturing geo-spatial relevance:**
  - Prior scores for regions (query and/or document scopes)
    - Area of encompassing region, population, ...
  - Distances over taxonomies of administrative place names
  - Approximate matches between place names
  - Overlap between geo-spatial regions (e.g., IoU)
  - Distance (e.g., between centroids or sets of points)

Neural Retrieval Methods



GIR Query Processing

- Documents are processed offline.
  - E.g., toponym resolution and document geo-coding.
- Indexes are also created offline.

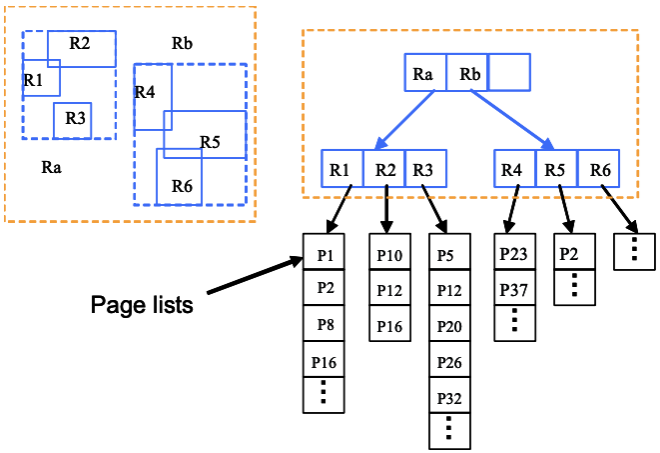
- Query is initially parsed into individual components.
- Index(es) used to filter candidate resources:
  - Satisfy the hard (spatial and/or thematic) restrictions;
  - Consider top-k pruning;
  - Retrieve information required for ranking.
- Ranking signals are computed for candidates.
- Final ranking is produced.

Spatial Data Structures for Indexing

- **Space-driven structures** using partitioning/mapping strategies to decompose 2D plane into a list of cells:
  - Fixed grid indexes sorting cells through space-filling curves.
  - Quadtrees, in which the resolution is varied according to the density of the spatial objects.
  - KD-trees, in which index nodes split the data based on coordinates along an interleaved axis.
  - HEALPix, geohash and similar (e.g., often using hashing).

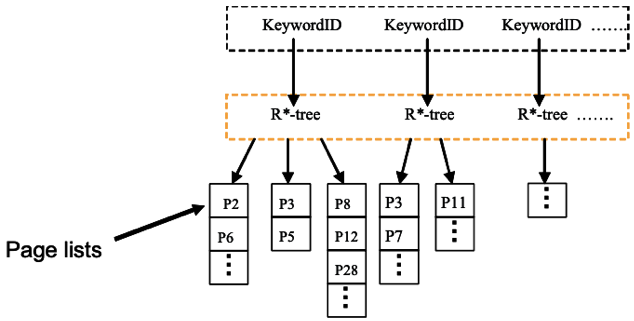
- **Data-driven structures** using the idea of spatial containment relationships instead of the order of the index:
  - R-trees consisting of hierarchical indexes on the minimum-bounding rectangles of geometries.
  - Many variants of the R-tree adding improvements to the construction process (R*-tree).

Combining Spatial Indexes with Inverted Lists

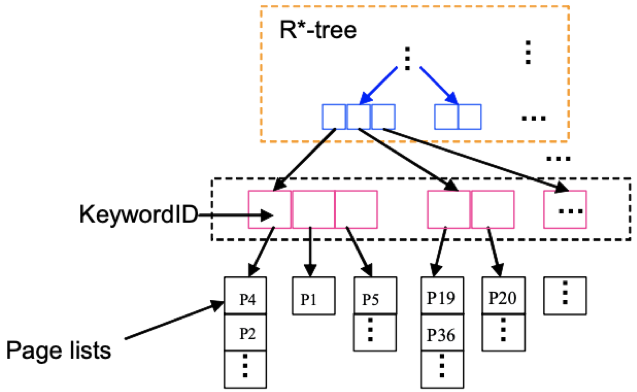


Multiple strategies in the literature:

- Handle region IDs within inverted index;
- Separate spatial and textual indexes;
- Hybrid index structures;
- Optimizations for top-k retrieval.



Primary Text Index



Primary Spatial Index

Combining Thematic and Spatial Similarity

Different options have been explored

- Filter resources according to hard geo-spatial restrictions, and keep text-based ranking scores.
- Filter resources according to key-word matches and rank by geo-spatial proximity or overlap.
- Linear combinations of thematic and geo-spatial signals, with weights defined by experts (*or controlled by the user*).
- Combinations of thematic and geo-spatial signals with weights inferred through machine learning.

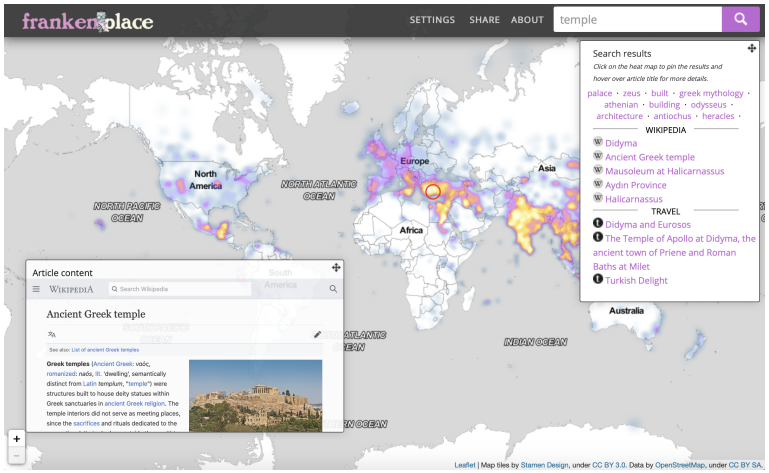
Other Ranking-Related Aspects

- Consider **spatial diversity** in ranked lists:
  - Results featuring many different (but spatially relevant) locations may satisfy users better.
  - Use diversity scores within ranking formula, giving higher weight to resources far away from previous ones.
  - Use algorithms aiming for high spatial coverage (e.g., interleaving results from different regions).

- **Cluster search results** according to geo-spatial regions:
  - Resources that are spatially near should behave similarly with respect to relevance to information needs.
  - Organize search results (e.g., administrative geography).

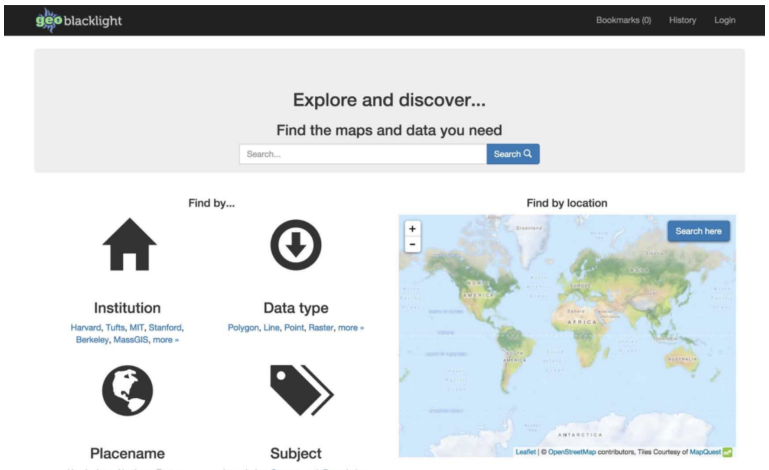
Putting It All Together

Exemplar Systems



www.frankenplace.com

newsstand.umiacs.umd.edu/



geoblacklight.org

emm.newsbrief.eu/overview.html

Evaluation Campaigns (1)

- Evaluation experiments require document collections for which there are a set of known information needs (i.e., topics) together with relevance judgements.
- Modern large-scale datasets such as MS-MARCO do contain examples of spatio-temporal information needs.

Joint evaluation campaign focused on GIR — GeoCLEF

- Task at CLEF between 2005 and 2008;
- Topics were increasingly complex (25 per edition);
- Retrieval from collections of newspaper articles;
- Different languages in either mono-lingual or multi-lingual scenarios;
- Geographically challenging topic set, with long descriptions for inf. needs;
- Binary relevance judgements collected through pooling;
- Dataset currently available through ELRA.
- Sub-tasks considered in 2007 (query parsing) and 2008 (GikiP).

Evaluation Campaigns (2)

GikiCLEF

- Pilot task in GeoCLEF 2008 and full-edition in CLEF 2009;
- Similar to question-answering (i.e., complex natural language questions);
- Identify Wikipedia articles answering geo-spatial needs (reasoning);
- Different languages in either mono-lingual or multi-lingual scenarios.

GeoTime

- Task associated to NTCIR in 2008 and 2009;
- Mixed spatial and temporal information needs;
- English, Japanese and Korean news corpora.

Future Challenges

- Integrate developments in text geo-parsing, document geo-coding, and spatial role labeling.
- Combining temporal and geographical retrieval.
- Improving multi-media, multi-modal and multi-lingual GIR.
- Combining neural methods with reasoning approaches.
  - Better support to geo-spatial question answering.
- Support text search over geo-spatial datasets.
  - Captioning remote sensing images.
  - Visual question answering over remote sensing data.

Questions?

Abstract

Text Meets Space: Geographic Content Extraction, Resolution and Information Retrieval

J.L. Leidner, B. Martins, K. McDonough and R.S. Purves

In this tutorial, we will review the basic concepts of, methods for, and applications of geographic information retrieval, also showing some possible applications in fields such as the digital humanities. The tutorial is organized in four parts. First we introduce some basic ideas about geography, and demonstrate why text is a powerful way of exploring relevant questions. We then introduce a basic end-to-end pipeline discussing geographic information in documents, spatial and multi-dimensional indexing, and spatial retrieval and spatial filtering. After showing a range of possible applications, we conclude with suggestions for future work in the area.

Zoom Instructions

https://zoom.us/j/100328143?pwd=cHFBbTZ1eVFpNFhwTXVnK0x1dCtRUT09

Meeting ID: 100 328 143
Password: 330739