Introduction

GIR systems typically consist of more than one index. This implementation so far uses two separate indexes, one for terms and one for spatial features, to calculate the relevance of a document to a query. What all these implementations have in common is the <Theme><Spatial Relationship><Location> triplet.

<Theme>/Topic

This input corresponds to the <theme> part in the triplet. Terms specified here (query terms) will undergo the same processing chain as indexed document terms so that the matching process can find the corresponding document terms in the index. Processing in this implementation includes:

- Stop word removal
- Lowercasing
- Stemming
- Tokenization

The following example shows the different steps applied on an input text.

1. Stop word removal & lower-casing

Before

Obschon der Piz d'Err 3378m neben dem Piz Calderas 3397m nur der zweithöchste Berg der Err-Gruppe ist, ist er der namengebende Hauptgipfel des Massiv.

After

obschon piz d'err 3378m neben piz calderas 3397m zweithöchste berg err gruppe namengebende hauptgipfel massiv

2. Stemming

Before

obschon piz d'err 3378m neben piz calderas 3397m zweithöchste berg err gruppe namengebende hauptgipfel massiv

After

obscho pix d'err 3378m neb pix caldera 3397m zweithochst berg err grupp namengeb hauptgipfel massiv

3. Tokenization

Before

obscho pix d'err 3378m neb pix caldera 3397m zweithochst berg err grupp namengeb hauptgipfel massiv

After

obscho	3378m	caldera	berg	namengeb
pix	neb	3397m	err	hauptgipfel
d'err	pix	zweithochst	grupp	massiv

The tokens resulting after the last step are then used to retrieve the documents containing those terms in the (inverted) index.

Term intersection

OR	Result document needs to contain **at least one** of the query terms to be considered relevant.
AND	Result document needs to contain **all** query terms to be considered relevant.

Text similarity

Implementation details can be found in [1]. This implementation provides several strategies to assess the similarity of documents which are explained briefly in the following table:

Boolean	If the term occurs in the document, the document is assigned the score 1.
TF-IDF	In this system, TF-IDF queries are the **sum of all tf-idf values of all the query terms** contained in a document. The following different formulas for calculating tf and idf are implemented: term_idf1 = $\log_2(N/n_i)$ term_idf2 = $\log_2(1+(N/n_i))$ doc_tf1 = freq doc_tf2 = 1+($\log_2$ (freq))
TF-IDF1	Sum over all query terms of (doc_tf1*term_idf1)
TF-IDF2	Sum over all query terms of (doc_tf2)
TF-IDF3	Sum over all query terms of (doc_tf2*term_idf1)
Cosine1	query weight: ((0.5*(freq*maxFreq))+0.5)*term_idf1 doc weight: doc_tf1*term_idf1
Cosine2	query weight: term_idf2 doc weight: doc_tf2
Cosine3	query weight: (1+($\log_2$ (freq)))*term_idf1 doc weight: doc_tf2*term_idf1
BM1	probabilistic model (only idf)
BM11	BM25 with b = 1
BM15	BM25 with b = 0
BM25	BM25 with k1 = 1.2 and b = 0.75. Doc length normalisation with avg. doc length as number of indexed words.

Text-spatial intersection

OR	query result may be relevant to text query, spatial query, or both (union)
AND	query result has to be relevant for both text and spatial query (intersection)

Combination type

Because spatial and textual indexes are separated, they retrieve a score for each dimension. To combine all scores of a document, different strategies may be applied. The following are implemented:

BordaCount	Definitions according to [2]. See also http://tinyurl.com/ozmastersthesis chapter 2.4.3.1.
CombMin	
CombMax	
CombSum	
CombAnz	
CombMNZ	

<Spatial Relationship>

Describes how the spatial relevance is estimated.

In	Points of documents (centroids if document footprint is a polygon) inside the query footprint (MBR retrieved through Yahoo PlaceMaker). Score is 1 if inside, 0 if outside the query footprint.
Directional (north, east, west, south)	According to [3].
near (linear decay)	According to [3], however, decay is circular the query footprints centroid (MBR from PlaceMaker). Radius is a function of the MBR's diagonal/2. points at the centroid get a score of 1, points at the radius' distance from the centroid get a score of 0. Linearly decreasing score from the centroid to the circle boundaries. See also http://tinyurl.com/ozmastersthesis chapter 2.4.2.2.1: LinearNear.

<Location>

Specifies the query footprint. Uses PlaceMaker. Retrieves an MBR as query footprint. To make sure the query location is found, specify the country (e.g. "Wallis" alone my not retrieve a query footprint. "Wallis, Schweiz" on the other hand will).

Query results representation

20 results are currently retrieved maximally to keep loading times low. The map shows a random query footprint for a document if the query was text-only. If the query was text-spatial or spatial only, the shown query footprint in the map corresponds to the footprint that gave the document its spatial score. "Score" in all cases represents the combined score if text-spatial was queried. "show" will show a document's full text, the number of indexed words and words in the document, the documents size in bytes, and most importantly for combined scores: the initial score of text and spatial queries.

Visual representations of indexes

"show indexes" will show the text index on the left side containing only a subset of the indexed terms. Currently, only terms that occur 12 – 20 times are retrieved to keep loading times low. The spatial index shows a Quad-tree containing all the centroids of document footprints in the index.

1. Baeza-Yates, R. and B. Ribeiro-Neto, *Modern Information Retrieval*. 2008: Addison-Wesley Publishing Company. 960.
2. Palacio, D., et al., *On the evaluation of Geographic Information Retrieval systems.* International Journal on Digital Libraries, 2010. **11**(2): p. 91-109.
3. Purves, R.S., et al., *The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet.* International Journal of Geographical Information Science, 2007. **21**(7): p. 717-745.