AGILE: GIScience Series, 6, 12, 2025. https://doi.org/10.5194/agile-giss-6-12-2025
Proceedings of the 28th AGILE Conference on Geographic Information Science, 10–13 June 2025.
Eds.: Auriol Degbelo, Serena Coetzee, Carsten Keßler, Monika Sester, Sabine Timpf, Lars Bernard
This contribution underwent peer review based on a full paper submission.

Mining Meaningful Facets in Spatial Information Retrieval with Spatial Relevance Feedback

Martin Werner [1]

[1]Technische Universität München, School of Engineering and Design, Professorship Big Geospatial Data Management, Munich, Germany

Correspondence: Martin Werner (martin.werner@tum.de)

Abstract. Information Retrieval is a set of techniques related to identifying and selecting documents from a very large collection of candidate documents based on their content. Traditionally, information retrieval is based on text documents and terms and various techniques for ranking the relevance of terms in documents. As an extension and to simplify the interaction of a user, however, techniques have been added enabling facet search. In this case, a search based on keywords or phrases is conducted. While doing this step, statistics on very specific low-rank properties of the documents are collected, e.g., price range, user ratings, color, manufacturer. This is then presented to the user together with search results in order to allow the user to filter or refine the search with respect to these queries. In this paper, we ask the question how meaningful facets can be computed for spatial databases and how this can be used to explore spatio-textual datasets exploiting such facets as an intuitive yet powerful information discovery mechanism beyond semantic categories. We show the feasibility of this approach on synthetic datasets, OpenStreetMap data, Wikipedia data, and social media data.

Submission Type. Algorithm; Software;

BoK Concepts. [AM10] Data mining; [AM2] Query operations and query languages

Keywords. Geospatial Information Retrieval; Social Media Analytics; Spatio-textual Search

1 Introduction

The rise of social networks, positioning, navigation, and ubiquitous Internet access has led to a situation in which an ever growing amount of textual information (e.g., social media messages, encyclopedia entries, messages, news articles) has become available for processing. At the same time, a lot of information related to human population is georeferenced, either explicitly by giving place or coordinate information, or implicitly as the information content of a data object has a local relevance. As a consequence, we are left with datasets in which a small amount of the data is explicitly georeferenced, while the majority of the available data are not. For these datasets, novel techniques are needed that combine developments from the text search community and the spatial computing community (Hu et al., 2022). With this paper, we present a new idea towards this direction.

Information retrieval is a family of techniques that finds a set of documents from a large set of candidate documents following a concept of relevance. Traditionally, specific document models have been employed in which a document is represented as a set of terms together with their frequencies. In this context, term refers to words in the simplest approach. However, it commonly refers to the word stem which is a certain representation of the set of all flections of a word. For example, “company” and “companies” are mapped to their word stem “compan”. Finally, terms can also refer to other information such as n-grams representing sequences of characters. In this setting, a common measure of importance for terms in documents is given by its relative frequency within the document. Roughly speaking, if a term appears more often, then it is more relevant to the document.

Now, the collection of documents, often called corpus, is transformed into a data structure for efficient lookup, most often into some sort of inverted index. In this case, the relation of a term being an element of a document is “inverted” in the sense that the index does not list the words in documents, but the documents for given words. By sorting both the set of terms and the set of documents for each term, quite efficient lookups can be made not only for documents that contain certain words, but as well as for documents that contain multiple words. During this approach to information retrieval, two basic query semantics can be distinguished wherein one model is called Boolean search. In this case, a query consists of an expression built from terms using relations AND, OR, and NOT combined with

parantheses for precedence modification. The other model is known as probabilistic ranking queries. In this case, the Information Retrieval system is supposed to rank documents based on some notion of relevance for the given query and enable the quick retrieval of the top k documents according to this ranking function.

In order to empower the user with additional features beyond formulating queries manually, three major lines of work have been established: query augmentation, relevance feedback, and facet search.

In *query augmentation* (Carpineto and Romano, 2012), the set of documents mined from a given query is used to extend the query with additional terms probably expanding across semantic relations, otherwise hidden. In this case, a query is used to construct a small result set. Then, the top terms from the result set can be carefully added to the query and the query is re-run. In fact, this means that words that are important to the top results of an initial search are used to broaden the coverage of the ranking function. This can be helpful to overcome barriers created by synonyms across languages. However, this scheme will also diverge from the initial meaning of the query such that one has to control the amount of terms being added and integrate user feedback in the search expansion in order not to accidentally reduce the specificity of the query beyond what is intended. Some authors present results on text datasets that even claim query expansion is often at least not positive for the power of the search system (Peat and Willett, 1991).

The second mechanism of *relevance feedback* is quite similar (Harman, 1992). The aim is to expand the query with more relevant information. But this time, the relevant information is obtained by presenting the user with a few results, either top results or maybe a sample, and asking the user to mark some of these results as either "good" or "bad" in terms of the query intention. Then, the Information Retrieval system can use the probabilistic model reversely to identify the terms which would have realized a minimal average score for the set of "bad" results and a maximal average score for the set of "good" results from the initial search. This refinement can be iterated.

The third technique is known as *faceting or facet search*. Facets are typically categorical properties that are associated with each document. For example, a hotel might be associated with the number of hotel stars or a shirt might be associated with all colors in which it is available. The most common case of using facets is to enable the user to quickly narrow down a search by proposing him the facet values of the initial research result (e.g., the price range or number of hotel stars) in order to filter the result set of the otherwise quite unspecific search. For example, searching for a hotel in a tourist city near the beach is likely not very specific as there will be thousands of offers. But presenting the user with semantic key information like price range, availability of services such as parking or restaurant, and distance to the beach, can help identify more relevant results that would be difficult to obtain by pure text search.

The main contributions of this paper in this context are:

- The introduction of a novel concept of supervised and unsupervised spatial facet search;
- The implementation of a scalable interactive system for spatial facet search;
- The demonstration of the power of a novel and interactive query mechanism in multiple, large real-world datasets.

In contrast to most related work, spatial facets are not applying spatial information directly within the search. Instead, we mine textual representations of the involved spatial objects. It is worth noting that the mechanism is designed to work on hybrid datasets where some, but not all of the documents, are georeferenced. This makes it significantly different from other related work in which a spatial component is prescribed for all documents. This is also the nature of common big datasets of text such as those mined from social media (e.g., from Twitter or X) or from Wikipedia wherein not all documents have geolocation attached.

The remainder of this paper is structured as follows: In Section 2, we present traditional facet search as part of information retrieval and introduce our novel spatial facet search queries in this context and give examples on synthetic data. Then, in Section 3, we explore the behavior on two real-world datasets including (1) a high-quality corpus of rather long documents mined from OpenStreetMap and German Wikipedia and, (2) a huge real-world dataset of noisy and low-quality short text taken from the Twitter public sample. Based on the insights of the case studies, we remark on open problems and research directions in Section 4. Section 5 gives links to related work in information retrieval and spatial keyword search. Finally, Section 6 concludes the paper.

2 Spatial Facet Search and Spatial Facet Mining

In this section, we describe some aspects of the class of information retrieval system underlying our implementation and experiments, then introduce the spatial facet queries.

2.1 Motivation and Setting

Most information retrieval systems have two basic operations. One is **indexing** which is bringing a given document into a searchable representation by organizing it inside an optimized data structure called index. The second one is **querying** in which the index is exploited to quickly scan candidates for search results, rank them based on a criterion of relevance and present them to the user.

For the indexing operation, keyword search engines are usually based on a data structure of an inverted index. These data structures are based on indexing strings by

managing ordered lists of document identifiers of all the elements containing a given keyword.

For the basic keyword search operation, each keyword of the query gives rise to an ordered list of document identifiers and thus, they can be iterated together in an ordered fashion by advancing the list pointer with minimal document ID. In this way, even negated terms can be easily tracked and documents that contain the given word can be removed.

These indices can be augmented to define a query extension mechanism. Given a set of document IDs, one can efficiently propose keywords to add to the query in order to maximize the rank of the set of given document IDs. In order to do this, a second inverted index is maintained in which all terms are indexed on documents. That is, we can easily retrieve all terms given in a document. This facility is useful for query augmentation as in the following definition.

Definition 1. *Given a database of documents D composed of documents d_i each consisting of weighted terms (t_j, ω_j), a keyword query Q, and a set of documents $E = \{e_i\}$, the **augmentation query for** (Q, E) **in** D returns a set of terms to add (OR or AND) to the query in order to maximize the rank of the elements of E. These keywords can be used to "augment" the query.*

One design consideration for such information retrieval systems is that the index files contain the right amount and selection of information. The documents themselves are usually held in some background storage and are only materialized on an explicit request, e.g., when presenting the top 10 results to the user. But, it is equally important to include additional information in the index itself such that it can be used in ranking and, for more complex query semantics to avoid unnecessary document materialization.

For example, **positional information** of terms (where in the text they are observed in relation to other keywords) allows for phrase queries and precedence queries giving all documents in which a word precedes another.

In addition to positional information, some information can be put into the index as well which helps the user interact with the search. A generic way is to assign a handful of arbitrary values to each of the documents. For example, we can add Boolean terms for metadata such as the filetype allowing for querying only documents in the PDF file format.

One common practice for embedding such information is by prefixing. For the sake of keyword search, text is often preprocessed (lower-cased, stemmed, cleaned). But if only lowercase letters appear, it is a natural choice to prefix terms with upper case letters and to form queries using these prefixes. For example, we could index a PDF file generating a set of terms contained in the PDF file. We prefix all of these terms with the letter 'Z'. We then add a term "Tpdf" where the prefix "T" is understood as filetype. Then a query like "keyword query words filetype:pdf" could be translated to a query "Zkeyword Zquery Zwords +Tpdf" actually searching for content words as given, but constraining the search to documents that have been marked as being a PDF file.

The downside of embedding data with prefixing is that the data itself is represented as a subset of the terms. For some data, this does not make too much sense, and therefore, additional values can be stored with each document in a special place that is materialized to the main memory on index traversal. At the same time, it is inefficient to look up keywords given a document ID in the secondary inverted (document=>terms) indices during index traversal in the primary index (term=>document) as this would lead to frequent cache misses, paging overhead, and random disk I/O. That is, we might want to embed certain values such as a file type specification directly into the primary index being traversed. In addition, certain values with complex semantics including timestamps or measures will need special treatment during search and could not even be covered by adding prefixed keywords. Hence, we need to be able to store a selected amount of special information together with our document IDs right within the primary inverted index.

One central technique that needs such non-keyword values is facet search. In traditional facet search, a query is run over a much larger range of results in the background and the values of the result are collected and presented to the user as options to narrow down the search.

Definition 2. *Given a database of documents D composed of documents d_i each consisting of weighted terms and a few values v_k, e.g.,*

$$d_i = (\{(t_j, \omega_j)\}, v_0, v_1, \ldots)$$

*and a query Q, the **facet query of depth** N **on a value slot** v_k **with aggregation function** C is defined to be the query that evaluates the N very well-ranked documents[1], reads their value slot v_i and maintains an aggregate of all visited documents using an online aggregation function C.*

In a web shop, value slots can be, for example, assigned to price and color. Then, the search engine visits a few thousand objects fulfilling the given query, counts the occurrences of values and presents the most frequent values as additional filters in a sidebar. In practice, for the values in such facets, suitably prefixed terms are added to the documents in order to run the facet-augmented queries efficiently.

2.2 Supervised Spatial Facet Search

Our approach to facets in spatio-textual databases is now to turn this faceting mechanism into a spatial search tool. First, we reserve a slot for the location of a document given

[1] For implementation efficiency, neither ordering nor strict top-k is enforced in this setting, instead the probabilistic bounds used for efficient index traversal introduce a small error.

as a pair of coordinates. Then, a given query is executed and we collect information from this value slot for a comparably large number of well-ranked documents resulting in a set of points.

Introducing spatial operations and relations on this set of points now defines spatial facet search:

Definition 3. *Given*

1. *a database D of documents d_i, some of which associated with spatial location p_i,*
2. *a keyword query Q,*
3. *a number N of documents to inspect,*
4. *and a spatial polygon P .*

*The **supervised spatial facet query***

1. *builds a set from the document IDs of the top N well-ranked documents regarding Q,*
2. *filters this set with respect to P,*
3. *computes augmentation terms using any query augmentation scheme extending Q to maximize the rank of the document in P.*

In this setting, the spatial predicate "being contained in P" can be replaced with more complex spatial predicates and the query augmentation scheme can involve the user through term proposal or "blindly" augment the query with the proposed terms.

If the documents from the selected spatial region can be characterized by a text query to some extent, this mechanism will propose suitable search terms to the user. While this is similar to a spatial filter in which only documents within the polygon P are being returned, it differs in two important aspects:

- First, by representing the spatial constraint as a textual representation, the search is still applicable for retrieval from non-spatial documents,
- Second, the user can observe the importance of a selected region for certain terms by observing the augmentation scheme for the same query in different spatial regions P.

In this context, the **supervised spatial facet** for a given tuple of a database, a query, a number of documents to inspect, and a spatial predicate are defined to be the set of proposed terms.

2.2.1 Example of a Supervised Spatial Facet Search Scenario

Consider Figure 1(a) which shows a domain X wherein documents are located as points with two colors. The red and green bullets represent documents "red green" and, similarly, the red and blue bullets represent documents of content "red blue". The figure further depicts three regions: U and V which cover similar documents and U^- which covers only a single document "red green".

Now, assume you are searching for the term "red". This term is contained in all depicted documents, therefore, all points are results of the same rank.

When running a supervised spatial facet query $(Q = \text{red}, U)$, the search for the keyword "red" will visit all documents. The supervision from the polygon U, however, reveals all documents of the type "red green". As a consequence, the information retrieval system will propose to add the keyword "green". That is, the augmented query will be $Q = \text{red OR green}$. Similarly, when faceting $(Q = \text{red}, V)$, the outcome would be the opposite. This means the system will propose to add the keyword "blue". However, if the polygon does contain green and blue in the same proportion, it won't propose a keyword for an extension. Finally, when faceting on $(Q = \text{red}, U^-)$, we will as well be proposed to add the keyword "green". The resulting query for "red OR green" would then be able to extend over the whole spatial region U.

In summary, when we create a facet out of all documents that lie within the set U or even the smaller set U^- that are discovered in a search, we will discover the word "green" and add it to the query.

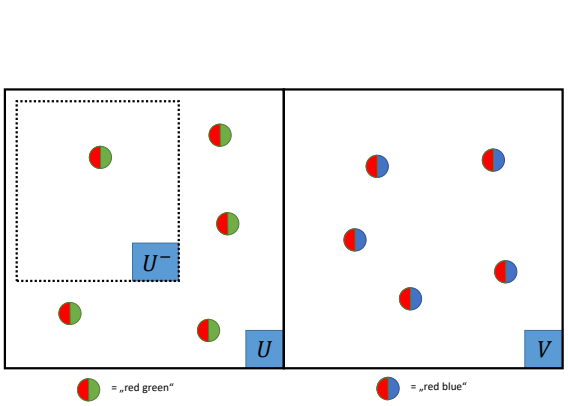
By running this augmented query (irrelevant of whether augmented on U or on U^- as weights are not relevant in this example), we will recover the set of documents within U in our search.

From an application point of view, this means that we can induce a spatial constraint on the query in this dataset, but neither by constraining document locations (result set filtering) nor by extending the information retrieval model (spatial information retrieval), but just by finding words that support our expected result set.

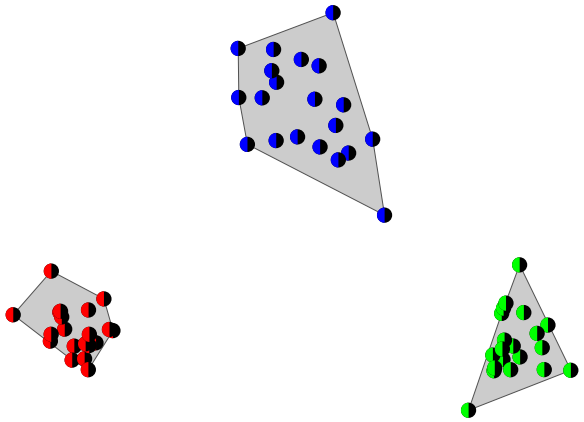
Note that even in this oversimplified example, the powerful aspect is that given a specific spatial object (e.g., U^-), a spatially extended result set is generated. In other words, queries can extrapolate spatial inputs. This is not true for traditional spatial information retrieval based on filtering or extending with spatial predicates.

Furthermore, note that the whole system is nowhere constrained with respect to connectivity. The polygon U could be split into a multipolygon of many pieces which would not affect the efficiency of the approach beyond a more involved point in a multipolygon test. This is in contrast to many spatio-textual indexing approaches in which space is represented by rectangles and query processing complexity is increasing with the number of rectangles. This means non-local facets, like facets for all major cities of a country, can be generated comparably easy.

On the negative side, however, note that we expect the spatial features to be expressible as a query. If this is not the



(a) A simple example of a dataset. To the left, within the set U, documents contain the two colors as terms, and similarly on the set V



(b) Result of Facet Clustering in the Unsupervised Facet Mining Example

Figure 1. Synthetic Datasets for Illustrating Spatial Facet Search

case, the system must fail. Or more concretely, we can only focus our search on spatial facets that are supported by words with a sufficiently local spatial distribution giving a limit to the applicability of this approach. In real-world spatio-textual datasets, such words are very common as local place names.

2.3 Unsupervised Spatial Facet Search

We introduce an unsupervised variant of spatial facet search by not relying on a given spatial predicate but rather, mining spatial specifications using clustering:

Definition 4. *Given*

1. *a database D of documents d_i, some of which associated with spatial location p_i,*
2. *a keyword query Q,*
3. *and a number N of documents to inspect.*

*The **unsupervised spatial facet query***

1. *builds a set from the document IDs of the top N well-ranked documents regarding Q,*
2. *clusters the point data using any spatial clustering algorithm*
3. *computes a spatial summary (e.g., convex hull, α-shape) of the relevant clusters.*

Thus, an unsupervised spatial facet query returns a set of spatial predicates for which supervised spatial facet queries seem reasonable. Such proposed regions can be ranked by the difference in the average rank of the documents in the region for the query Q and the augmented query $\tilde{Q}$ documents, simply speaking by their "augmentability".

Note that the probabilistic weighting of the documents identified could be integrated into the spatial clustering, but we leave this as an area for future work.

We introduce a final query in this context which completes our spatial information retrieval system in the following sense: while the supervised spatial facet query takes spatial knowledge to steer a search and the unsupervised spatial facet query mines spatial knowledge from the search, the spatial response map query mines spatial information from the behaviour of text search.

Definition 5. *Given a database D of documents d_i, a spatial subdivision of space (e.g., a grid) and a term t, the **spatial response map of** t is defined to be the map coloured from the weight of the term t for augmenting the query Q towards the current region of the grid. If the term is not proposed by a spatial supervised facet query at all in a specific region, the weight is considered a missing value.*

2.3.1 Example Unsupervised Spatial Facet Query

For illustrating the unsupervised spatial facet mining approach, we create a simple clustered dataset similar to the dataset before. We create a dataset of spatial documents from three bivariate normal distributions $\mathcal{N}_2$ with means and covariances as follows:

$$U_1 = \mathcal{N}_2\left(\mu = \begin{pmatrix}-10\\0\end{pmatrix}, \Sigma = \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right), \text{“black red”},$$
$$U_2 = \mathcal{N}_2\left(\mu = \begin{pmatrix}10\\0\end{pmatrix}, \Sigma = \begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}\right), \text{“black green”},$$
$$U_3 = \mathcal{N}_2\left(\mu = \begin{pmatrix}0\\10\end{pmatrix}, \Sigma = \begin{pmatrix}3 & 0\\0 & 3\end{pmatrix}\right), \text{“black blue”},$$
$$X = U_1 \cup U_2 \cup U_3$$

This dataset is used for querying “black”, which is contained and relevant for all documents. Therefore, the spatial values observed consist of all points. We run a clustering algorithm (DBScan with $\epsilon = 5, \text{minPts} = 3$, Ester et al. (1996)) on top of this and get the three clusters colored in Figure 1(b).

For each of these clusters, the search engine is asked for keywords maximizing the rank of the documents within these clusters and the engine reveals, as expected, that the most helpful words are “red” for the first cluster (bottom-left, documents of type “black red”), “green” for the second cluster (bottom-right, documents of type “black green”), and “blue” for the third cluster (top-middle, documents of type “black blue”)[2].

2.4 Spatial Information Retrieval Process

With this paper, we extend the information retrieval process with the following queries exploiting spatial information assigned with documents:

- The supervised spatial facet query mining the most relevant keywords in a given polygon (Def. 3),
- The unsupervised spatial facet query clustering the location of high-ranked documents providing both spatial summaries (e.g., polygons) and terms characterizing spatial clusters (Def. 4),
- The response map query in which the spatial relevance of a given keyword for a given query is visualized (Def. 5).

Based on the availability of these queries, the information retrieval process is refined as depicted in Figure 2. One starts with a search based on a given basic query input by the user. As usual in the domain of keyword-based search, this query can consist of ranking terms, Boolean terms, temporal range expressions, negations, wildcards, quotations to avoid stemming and use positional information.

The user now enters a query. For example, “Rathaus + "München"”. This query searches for documents, where Rathaus (German word for town hall) is relevant and which contain (Boolean search) the unstemmed term “München” (German spelling of Munich). Putting “München” in quotation marks is important as the stemming mechanism would otherwise turn “München” into “Munch” which refers to a widely known Norwegian painter.

This search is conducted and while searching for the most relevant results, we collect location and score information for all documents visited in this search.

In a second step, we visualize this spatial information mined during result set generation as depicted in Figure 4. In this visualization, a spatial selection takes place either by user input or by clustering. These selected sets of documents can be used for steering the search towards this set or away from this set by negating the augmentation expression. With this spatial selection in place, we compute the most meaningful terms from the involved documents separating them from the rest of the database. That is, a set of terms that would increase the ranking of the selected documents while decreasing the rank of others.

Finally, some of these augmentation sets can be added or added with negation to the query refining the search. As the result of this process is a valid query, we can iterate this process in a loop until a convincing result is generated.

2.5 Data and Software Availability

In general, data and code for this publication are available according to the following details:

The demonstrative implementation is available from https://github.com/tum-bgd/spatial_facet_search and is licensed under Apache 2.0 license (our code). Note that other open-source licenses can apply to dependencies. The implementation supports all queries.

Due to legal constraints and privacy concerns, we do not share the dataset based on Twitter data; the dataset derived from OpenStreetMap, however, is available at https://api.bgd.ed.tum.de/datasets/spatialfacetsearch.

3 Case Studies

In this section, we present some case studies on real-world datasets in order to show the feasibility of this approach.

3.1 Case Study on Wikipedia and OpenStreetMap

For experimental validation, we created a spatial subset of German Wikipedia. The OpenStreetMap project allows for connecting OSM spatial objects like houses, points, streets, or polygons with Wikipedia articles. This feature has good coverage in Germany.

We joined both datasets using this link. For each object in a Germany extract of OpenStreetMap that contains a link to Wikipedia, we extracted both the spatial information from OpenStreetMap and the textual information from

[2]The two synthetic experiments can be run from SimpleDatasets.ipynb in the published source code repository https://www.github.com/tumbgd/spatial_facet_search

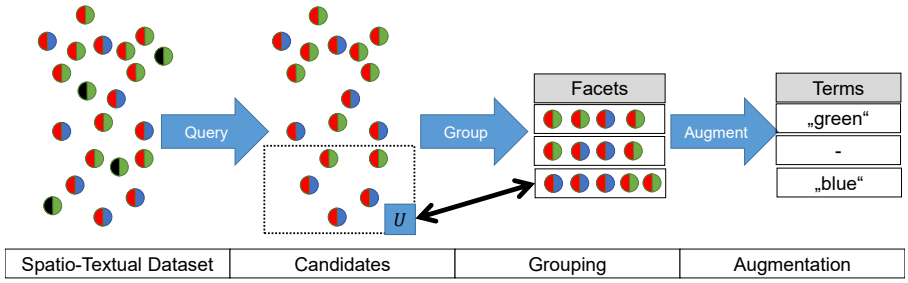


Figure 2. Spatial Information Retrieval Process with Augmentation

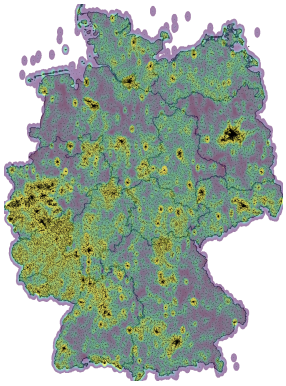


Figure 3. The spatial distribution of the dataset integrating OpenStreetMap references and German Wikipedia articles

Wikipedia including an explicit title and slightly cleaned text.

The dataset contains about 72,500 different spatial objects. But as in some cases where more than one object refers to the same Wikipedia page, our decision to index on Wikipedia ID instead of OSM ID reduced the dataset to an index of 56,287 documents (average document length: 1507.8 characters, 2,079,264 distinct terms). Figure 3 depicts the spatial distribution of this corpus.

In this example, we search for the term “Kirche” (church) in the dataset. More concretely, we will search for church and observe the set of documents touched during search generating a spatial facet. This contains all documents together with their score that have been visited while traversing the inverted index for the given query “church”. Now, we spatially select some documents and compute the terms for query augmentation that maximize the rank of selected documents. Therefore, we use two datasets on different scales: German states (“Bundesland”) as well as counties (e.g “Kreise und Regierungsbezirke”).

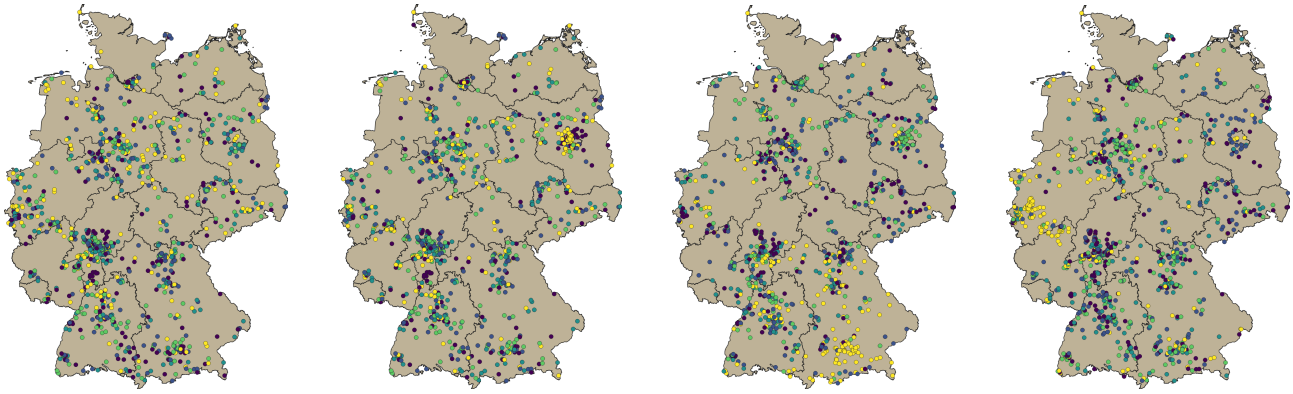
Figure 4 shows the results of augmenting the search for the counties Berlin, Munich and Rhein-Kreis Neuss. We show both the spatial facet of searching for “Kirche” (church), and the spatially augmented results from adding the top ten terms to the ranking search. Obviously, adding the self-discovered terms leads to a good focus on the selected region without filtering away all documents that did not exactly fit the spatial region. This is especially important for Berlin where the size of the spatial selection is significantly smaller than the extent of the urban agglomeration around the city.

We can as well look at the terms and their weight that have been discovered through this spatial selection. In the case of Munich, we discover “München” with a score of 56.8, “Bürgermeister” (“mayor”) with score 38.77, “Bahn” (“train”) with score 34.0 and “oberbayerisch” (“Upper Bavarian”) with score 33.1. The first term “München” is obviously a toponym. The terms mayor and train are slightly surprising, and Upper Bavarian is again a clear and nice toponym. The surprising terms are query- and corpus-specific toponyms. That is, as soon as we are already searching for Munich and church, giving some politicians roles as keywords, is helpful within Wikipedia.

For the case of Berlin, we get some religious terms first: kirch (21.5) which is a stem from Kirche meaning church, Nazarene (17.2), jesus (16.16), and followed by very local toponyms like Schöneberg (15.5) being a district of the city of Berlin.

For Rhein-Kreis Neuss which is a large county near Cologne, we reveal keywords “Neuss” (62.9), “Köln” (“Cologne”, 36.06), “Grevenbroich” (30.79), “Meerbusch” (30.67) and “Büderich” (30.49) for augmentation. This is interesting as this is a representative set of cities from this area. Only Cologne is not part of Rhein-Kreis Neuss though these counties touch with each other. But Cologne is the archdiocese for all churches in Rhein-Kreis Neuss explaining the importance of this city for churches in the area and the relevance of this city for the county within Wikipedia.

These examples show that the proposed process of spatially focusing to a region, not by means of integrating exact spatial search with probabilistic information retrieval, but rather by finding a probabilistic description of the spatial region which can be combined inside the same framework of probabilistic ranking queries is working as expected and beyond what spatial keyword search delivers. This is, however, due to the nature of the corpus being an encyclopedia: a spatially-referenced article is a long and detailed description of this very location providing a textual keyword link to relevant areas outside focus.



(a) The initial facet when searching for "Kirche" (church)

(b) Berlin (augmented)

(c) München (augmented)

(d) Rhein-Kreis Neuss (augmented)

Figure 4. The original and augmented facets for three spatial areas in Germany. The yellow dots represent top 20% records according to their score. One can clearly see how a yellow cluster is formed in the area of Berlin (North-East), the area of Munich (South-East), and near the river Rhine (Middle-West).

We also want to show examples where this approach fails. For example, when the spatial focus is not small enough such that no toponyms exist that correctly and sufficiently describe the spatially selected set of results. For this analysis, we used the German states in order to assess how far such large-scale spatial selections can be represented through query augmentation.

As you can see in Figure 5 for a few examples, this seems not to work. While augmenting changes in the score distribution as well as the set that is being visited by the search algorithm, one cannot easily see a spatial pattern in all of those examples with the exception of Sachsen.

For Sachsen, it is interesting that Sachsen and Sachsen-Anhalt are now covered by yellow dots indicating a significant increase in hits in this area. Looking into the augmenting word gives us an explanation. For North-Rhine Westfalia, the search has been extended with a stem "kirch" of "Kirche" (church) and stop words only (von [of], ist [is], ein [a or one], im [in], die [the]). These words will appear in all documents of Wikipedia as they belong to the core of the German language. In a certain sense, the query augmentation has modelled noise and finding out that these terms have higher scores in the documents found in North-Rhine Westfalia as opposed to the general.

For Sachsen, as well as "kirch" stemmed from "Kirche" meaning church, has been added and stop words like ("dem, wurde, ein") and words obviously irrelevant to the task ("ort" [place], "jahr" [year]). But, the word "Sachsen" itself has been identified leading to the fact that a significantly higher score is given to hits in "Saxen" (Saxony) and Sachsen-Anhalt (e.g., Lower Saxony). In this case, the toponym Sachsen that is somehow relevant to Sachsen has been correctly identified though it also applies to documents related to Sachsen-Anhalt (Lower Saxony).

In summary, we can conclude that injecting spatial information on a county-scale does not work in general but, if there exist widely-used toponyms at this scale, they will most likely get identified.

3.2 Experiments on a Twitter social media corpus

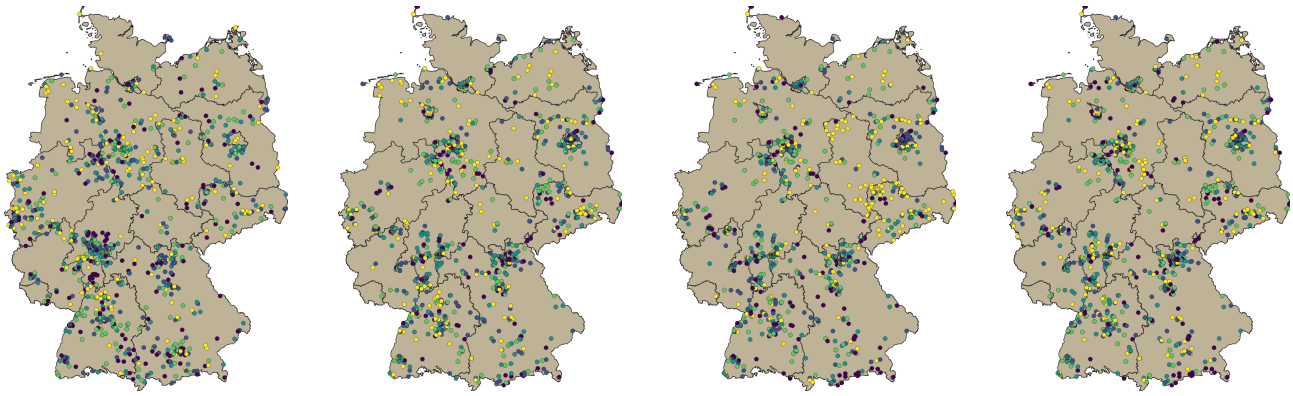
In this section, we analyze a sample of the Twitter social network API observed during 2017. We first index 13 million tweets (12,940,000) with georeference for a database with an average document length of 28.12 words.

As a first step, we implement an interactive search tool in a web browser in which we can define polygons and search for terms getting either the top results or the top augmentation queries for a query polygon that can be interactively edited.

Figure 6 depicts selected queries performed with the interactive web browser interface and their results. Within the given Twitter dataset, we first search for the term "apple" which refers to a fruit and a well-known computer company at the same time.

When running this query with faceting regions on very large scales, we are proposed with keywords like "iphone", "iphone x", and "tim cook". This is within our expectations as in these scales, the computer company Apple clearly dominates tweets. However, when faceting East of Sacramento, we also see proposals for "applehill" getting quite strong. This is a local association of farmers[3]. Furthermore, we can ask for the spatial response of the keyword "iphone" in relation to the query "apple". Figure 6(e) and Figure 6(f) depict the spatial response of this keyword in the top 20 query augmentation terms. The more red a rectangle is displayed, the stronger the weight of the term "iphone" in this area for faceting to this area. The interpretation of such maps actually depends on the context, for the term "iphone" on a global scale, we see urbanization

[3] https://applehill.com/



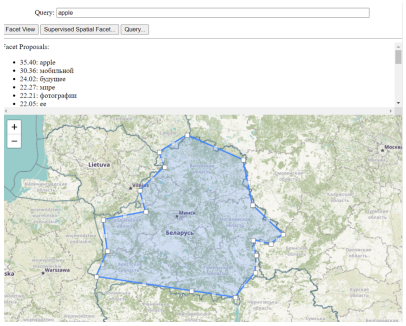
(a) The initial facet when searching for "Kirche" (church)

(b) North-Rhine Westfalia (augmented)

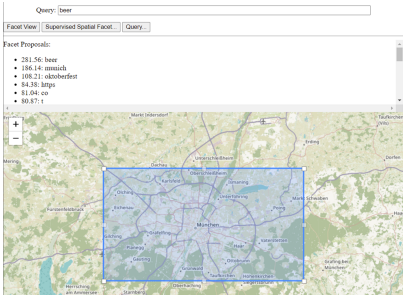
(c) Sachsen (Saxony)

(d) Bayern (Bavaria)

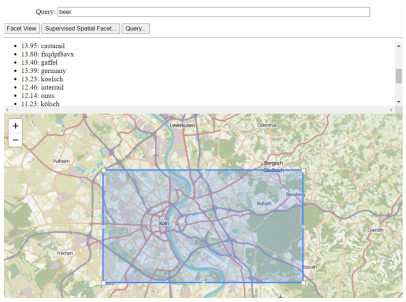
Figure 5. The original and augmented facets for three states in Germany. No clear spatial trend is visible with the notable exception of Sachsen (Saxony, Lower Saxony) giving rise to a yellow cluster in the Middle-East of Germany.



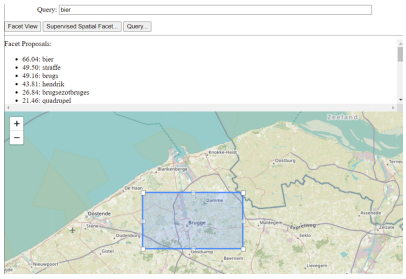
(a) Faceting "apple" over Belarus finds Cyrillic words. A nice example of the power of query augmentation.



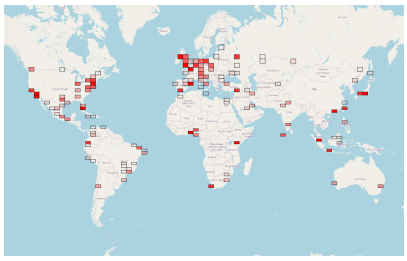
(b) Faceting "beer" over Munich reveals Oktoberfest (The world's largest beer festival).



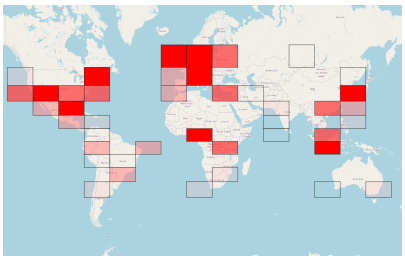
(c) Faceting "beer" over Cologne reveals Gaffel and Kölsch (Gaffel is a famous beer from around cologne and the type of beer is known as "Kölsch").



(d) Faceting "Bier" over Brugge reveals "Straffe Hendrik" (a famous beer from Brugge, see https://www.straffehendrik.be/en/home).



(e) A High-resolution Response Map for the keyword "iphone" in relation to a query "apple".



(f) A low-resolution response map for the keyword "iphone" in relation to a query "apple".

Figure 6. Spatial Facet Examples on Twitter Real-World Datasets

patterns (more coastal areas, strong Europe, strong China) showing that a meaningful, expected spatial feature was revealed.

Finally, consider Figure 6(a) in which we facet the query "apple" roughly to Belarus. In this case, it is nice to see that we discover words in Cyrillic script for mobile phone and photography. This could be used to conduct further searches in documents in Cyrillic script.

A second experiment on this real-world social media data reveals the power of the approach on the topic of beer. When faceting beer over Munich, we are left with proposals of "munich", "oktoberfest", "festival", "hellomunich", "sausage", "hofbräuhaus", "knuckle", and "german" (and some noise terms), see 6(b). This is a very reasonable breakdown of the Munich beer culture as reported by

tourists and local newspapers (e.g., “hellomunich”) in Munich.

Moving the window to Cologne brings up “köln”, “gaffel”, “koelsch”, “instabeer”, “colonia”, and “dom”, among other keywords. This, as well, is a good breakdown of tourist messaging including beer types and traditional words. See Figure6(c).

When moving the same frame over to Brugge (a Flemish city in Belgium), see Figure 6(d), and adjusting the search word to “Bier” which better captures the French and German word (“bière”, “Bier”), we discover keywords “straffe”, “brugs”, “hendrik”, “brugsezotbruges”, “quadrupel”, “tripel”, “belgium”, “smelling”, “admiring”, “yeast” and “gezelligheid”. All these words are related to the beer culture around Brugge. For example, the oldest brewery of the town, “Halve Maan” (founded 1546), is discovered together with their beers “Brugse Zot” and “Straffe Hendrik”. Furthermore, quadrupel and tripel are specific types of Belgian strong ales.

This experiment shows the remarkable exploration power of spatial facet searching and also, that tourists social media activity are a good measure of local cultural heritage like the history of beer brewing and drinking in Western Europe.

3.3 Discussion

We have presented two series of experiments on spatial facet search. In the first series, we took very high-quality, long text from the German Wikipedia with georeferences mined from OpenStreetMap. It revealed that spatial facet search in such clean data is very powerful and allowed us to discover relevant keywords and to extrapolate from a given spatial region. For example, identifying Cologne as the kernel and connection of many smaller cities around the river Rhine.

In addition, we showed that very weak text (Twitter without advanced pre-processing) with very limited availability of spatial information could successfully be mined spatially using our proposed framework.

Hence, we have shown that spatial facet search following the approach of this paper is a promising candidate for practical geospatial applications. The real value of the method, however, can only be evaluated in domain-specific research in the next years.

4 Remarks on Open Questions and Future Work

With this paper, we extend the literature with an example and process how modern information retrieval techniques such as facet search and relevance feedback can be combined in a spatial information retrieval setting. Though not covered in this article, it is obvious that the same approach can be used for temporal and spatio-temporal data by extending the polygon to a polytope in the space-time cube and the clustering from spatial to spatio-temporal. In practice, timestamps can be added in an additional value slot to allow for arbitrary influence on weighting.

In this section, we highlight a few open issues and areas of research which we think are both interesting and promising as follows:

Determine the Number of Documents for Facet Generation

One central parameter for this work is the number of documents that are visited during search. As a minimum, we will visit those documents that allow us to proof that we have found the top-k elements of the query. As a maximum, we could explore all documents. The first one is most likely too small as facets are only generated from as few results as are going to be displayed on screen. The last one is maybe too heavy reducing interactivity.

Incremental Weighted Stream Clustering

An obvious candidate technique for scale detection and automatic spatial augmentation is to apply unsupervised clustering methods combined with cluster quality observations. In this work, we relied on clustering approaches that need spatially local access to all points (e.g., k-means, DBScan, OPTICS). It is interesting to research how such algorithms can be replaced with incremental ones. Furthermore, the weight relative to the query should be incorporated into the query.

Toponym Quantification and Spatial Stop Words

We have seen in the example that some of the spatial query augmentations revealed toponyms and some revealed stop words. One interesting option is to query for all terms indexed in the database and quantify the spatial distribution. If it is sufficiently local - whatever this means - we have detected toponyms. If the spatial distributions remain comparable, however, we have clearly identified words without spatial meaning which might or might not be stop words. Such classification could be interesting both for exploring the corpus as well as for text mining on this corpus.

5 Related Work

This paper combines two major lines of research and development in computer science. First of all, it is related to keyword search, which is at the heart of the research field **information retrieval**. At the same time, it is related to spatial computing, most notably with papers related to **spatial keyword search** in which spatial and textual aspects are combined.

Information Retrieval (IR) is a traditional research topic in computing providing the backbone of the search engine-

empowered Internet. The basic model of information retrieval is a document model in which a document is a set of terms. Terms are scored relative to the document itself and to the overall set of documents trading of the frequency of the word in a document with the expected frequency of a word. One widely-known and highly traditional approach for this weighting is based on TF-IDF where the frequency of words within a document is related to the inverse document frequency capturing aspects of the fraction of documents that contain this word Jones (1972); Church and Gale (1999). This balances between so-called stop words like "and", "the", and "for" appearing frequently in all documents, and corpus-specific stop words like "References" when indexing scientific papers and the assumption that frequent words are more important than infrequent ones.

In our paper, however, we rely on the BM25 weighting scheme Robertson et al. (1999). A nice overview of common weighting schemes is given in Cummins and O'Riordan (2006).

Spatial data has attracted interest in the information retrieval community quite early. For example, the paper of Larson dating back to 1996 presents a nice introduction to spatial information retrieval, especially in putting it in contrast to exact database search Larson (1996). They already proposed a spatial browsing application bridging between term-based information retrieval and spatial querying. A more recent paper of Hariharan et al. discusses how spatial and textual information can be combined in a geographic information retrieval system: by means of separate indices as well as by hybrid indexing structures Hariharan et al. (2007).

This paper links the topic of this work to the domain of spatial computing and especially to a track of work related to **spatial keyword search**. A recent survey provides in-depth discussion of this field Chen et al. (2020). In this survey, the authors distinguish three types of queries: Boolean-Boolean (BB) in which spatial and textual queries are exact, Boolean-Ranking (BR) queries in which the spatial proximity is used for ranking (e.g., a top-kNN set of documents matching the query), and a Full Ranking Query in which a monotonous combination of a spatial and a textual relevance measure is used. Examples and related work for these three directions are given in the survey.

It is worth noting that these three query semantics do not fit our work. We differ in the fact that we use a purely textual ranking wherein spatial information is not used during ranking. But in order to make this spatial-aware, we encode the spatial nature in toponyms found in the corpus. Therefore, we represent a novel and fourth class of queries. We term this as Proposal-Ranking queries in the sense that the spatial information is used for keyword proposal and query augmentation while the ranking remains purely textual with the described advantages (covers non-spatial elements as well, can detect informative corpus-specific toponyms, remains fully explainable, avoids the tradeoff between spatial and textual aspects) and disadvantages (if such spatial toponyms don't exist, the search cannot be focused and the spatial information remains ineffective).

Recent work trying to combine information retrieval and deep learning-based natural language processing is an interesting development. This presents an interesting area of future research, especially in how geometric representations of text embeddings can be used to propose keywords or to induce an artificial geometry for spatial faceting within this geometry. Some initial work on spatial keyword search combined with text embedding is given in Oh et al. (2018) and the behavior of such models on Twitter has been discussed in Häberle et al. (2019).

6 Conclusion

With this paper, we introduced the novel problem of facet extraction from spatial information retrieval systems and proposed a few solutions to this which are useful in varying scenarios. We showed how the visiting of high-rank search results' associated spatial information can be used to provide advanced query augmentation interaction possibilities to users. It also provided that spatial search in the probabilistic document model is sometimes possible in practice.

This is an important aspect as many other papers try to extend information retrieval systems with exact spatial predicates which we think are not ultimately useful given the underlying assumption that **all** meaning of a document is encoded in the relations between terms, documents, and the corpus. With this assumption, a spatial search for documents is only sensible if there are keywords with a clear spatial focus. We show a mechanism, how such words can be found.

In contrast to many approaches from the deep learning and text mining field coming up today, it is worth noting that the whole system is explainable by design. As we encode all information retrieval parameters in a set of weighted terms and as we modify the terms only with user interaction, with which a user of the system is able to explain the search results. In practice, this means that words that might be subject to bias (including toponyms, gender, race, etc.) could be proposed by the algorithm. However, we can expect the user to be able to see this problem and to, probably, remove these problematic terms.

In summary, we propose a novel paradigm of information retrieval in spatial datasets that do not require **all** documents being spatial. Instead, we rely on the expectation that the interesting spatial aspects can be covered by keyword search. We provide a highly efficient open-source search engine including our techniques with a web user interface for easy adoption.

References

Carpineto, C. and Romano, G.: A survey of automatic query expansion in information retrieval, Acm Computing Surveys (CSUR), 44, 1–50, 2012.

Chen, L., Shang, S., Yang, C., and Li, J.: Spatial keyword search: a survey, GeoInformatica, 24, 85–106, 2020.

Church, K. and Gale, W.: Inverse document frequency (idf): A measure of deviations from poisson, in: Natural language processing using very large corpora, pp. 283–295, Springer, 1999.

Cummins, R. and O'Riordan, C.: Evolving local and global weighting schemes in information retrieval, Information Retrieval, 9, 311–330, 2006.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise., in: Kdd, vol. 96, pp. 226–231, 1996.

Hariharan, R., Hore, B., Li, C., and Mehrotra, S.: Processing spatial-keyword (SK) queries in geographic information retrieval (GIR) systems, in: 19th International Conference on Scientific and Statistical Database Management (SSDBM 2007), p. 16, IEEE, 2007.

Harman, D.: Relevance feedback revisited, in: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 1–10, 1992.

Hu, X., Zhou, Z., Li, H., Hu, Y., Gu, F., Kersten, J., Fan, H., and Klan, F.: Location reference recognition from texts: A survey and comparison, arXiv preprint arXiv:2207.01683, 2022.

Häberle, M., Werner, M., and Zhu, X.: Geo-Spatial Text-Mining from Twitter—A Feature Space Analysis with a view towards Building Classification in Urban Regions, European Journal of Remote Sensing, 52, 2–11, https://doi.org/10.1080/22797254.2019.1586451, 2019.

Jones, K. S.: A statistical interpretation of term specificity and its application in retrieval, Journal of documentation, 1972.

Larson, R. R.: Geographic information retrieval and spatial browsing, Geographic information systems and libraries: patrons, maps, and spatial information [papers presented at the 1995 Clinic on Library Applications of Data Processing, April 10-12, 1995], 1996.

Oh, S., Jung, H., Koo, J., and Kim, U.-M.: Efficient Method for Processing Range Spatial Keyword Queries Over Moving Objects Based on Word2Vec, in: International Conference on Human Interface and the Management of Information, pp. 620–639, Springer, 2018.

Peat, H. J. and Willett, P.: The limitations of term co-occurrence data for query expansion in document retrieval systems, Journal of the american society for information science, 42, 378–383, 1991.

Robertson, S. E., Walker, S., Beaulieu, M., and Willett, P.: Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track, Nist Special Publication SP, pp. 253–264, 1999.