The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing

Christopher B. Jones, Alia I. Abdelmoty, David Finch,
Gaihua Fu, and Subodh Vaid

School of Computer Science, Cardiff University, Cardiff, Wales, UK
{c.b.jones,a.i.abdelmoty,d.finch,gaihua.fu,s.vaid}@cs.cf.ac.uk

Abstract. The SPIRIT search engine provides a test bed for the development of web search technology that is specialised for access to geographical information. Major components include the user interface, a geographical ontology, maintenance and retrieval functions for a test collection of web documents, textual and spatial indexes, relevance ranking and metadata extraction. Here we summarise the functionality and interaction between these components before focusing on the design of the geo-ontology and the development of spatio-textual indexing methods. The geo-ontology supports functionality for disambiguation, query expansion, relevance ranking and metadata extraction. Geographical place names are accompanied by multiple geometric footprints and qualitative spatial relationships. Spatial indexing of documents has been integrated with text indexing through the use of spatio-textual keys in which terms are concatenated with spatial cells to which they relate. Preliminary experiments demonstrate considerable performance benefits when compared with pure text indexing and with text indexing followed by a spatial filtering stage.

1 Introduction

All aspects of human activity are rooted in geographic space in some respect. As a consequence many types of documents include references to geographical context, typically by means of place names. This common occurrence of geographical references is reflected in documents stored and retrieved on the world-wide web. If users of a web search engine wishes to find resources in which the subject matter is related to a particular place, then they can include the name of the place of interest in the search engine query. Conventional search engines treat the query place name in the same way as any other keyword and will retrieve documents that include the specified name. For some purposes this may be adequate, but there are many situations in which the user is interested in documents that relate to the same region of space as that specified by the place name, but which might not actually include the place name. This could occur if there were documents that used alternative names, or referred to places that were in or nearby the specified place. The process of exact match will also inevitably result in the retrieval of irrelevant documents due to multiple use of certain names to refer to

M.J. Egenhofer, C. Freksa, and H.J. Miller (Eds.): GIScience 2004, LNCS 3234, pp. 125–139, 2004.

different places, and the fact that place names often occur in the names of products or organisations that are not associated with the named place. There is a need therefore for spatially-aware search engines that can interpret the presence of a place in a query in an intelligent manner that results in improved quality of information retrieval.

The introduction of spatial awareness in a search engine poses several significant challenges. At the user interface it should be possible to recognize the presence of place names in a query expression. This implies the existence of a directory of place names, such as is found in gazetteers [Ale]. If the search engine is to be entirely general purpose it would need to maintain knowledge of the names of every place on Earth. Users may wish to qualify the spatial aspect of their search using spatial relationships, in which case it becomes necessary to interpret such relationships and transform them to the representation of an appropriate geographical region, or "***query footprint***" that can be used for search purposes. Here we refer to the knowledgebase required for these purposes as a geographical ontology, or geo-ontology. Retrieval of relevant documents requires a process of geographical metadata extraction whereby the geographical context of a document is determined by some form of analysis of the text. Techniques are needed for disambiguation of place names and for establishing the likelihood that the document's information content is actually related to particular names that are present. The geographical coverage of a document is referred to here as the "***document footprint***", which could consist of multiple parts. Once document footprints have been established, the possibility then arises for spatial indexing of the documents to facilitate fast access to documents pertaining to a given query footprint. Retrieval of documents must be followed by relevance ranking with respect both to their closeness to the query footprint and to thematic, possibly non-spatial, terms that the user has employed in the original query.

Recently a few search engines that are specialised with respect to geographic space have appeared (e.g. the vicinity products in the Mapblast[Vic] and Northern Lights web sites [Nor]; Mirago.com [Mir]; the experimental Google locational search engine [Goo]). Relatively little has been published on the technology that underlies these search engines, though there have been some published accounts of research efforts relating to particular aspects of the development of a geographical search engine functionality [DGS00,BCGM$^+$99,McC01]. In this paper we describe the architecture of the SPIRIT prototype geographical search engine that is currently under development and which is intended to address the challenges to spatial search that are listed above. The main components include a user interface, a test collection of web documents with associated search engine maintenance and retrieval functionality, a geographical ontology, relevance ranking procedures, and document metadata creation and enrichment procedures. The role of each of these components and the interactions between them are introduced, before focusing in more detail on two specific aspects of the search engine, concerning the geographical ontology, and integration of spatial indexing with text indexing. Future articles will focus on other components of the search engine.

In what follows we summarise related work on geographical search engines, before describing the overall architecture of the SPIRIT search engine in section 3. The design of the geographical ontology is then presented in section 4 along with a summary of the access methods that have been implemented for it, in order to support processes of document annotation and metadata extraction, user interaction and relevance ranking. In section 5 we describe a novel approach to combining text indexing with spatial indexing, using spatio-textual keys, and provide a preliminary experimental evaluation of the technique. The paper concludes by highlighting current research issues and summarising future work.

2 Related Work

Geographical web search facilities developed by the company Vicinity [Vic] are present on the Mapblast [Map] web site and the Northern Light search engine [Nor]. These facilities allow the user to enter part or all of an address in the USA or Canada, along with a category of interest and a search radius in miles. It appears that the tool translates the address to a map coordinate and expands the search to include other places within the specified radius, with the aid of a digital map. The Mirago search engine [Mir] provides what is referred to as a regional web search facility with a focus in the four countries of UK, Germany, France and Spain. Here the user can select individual regions of a country on which to focus their web search. Google have recently introduced a demonstration of locational web search based in the USA [Goo]. Like the Vicinity search tools it allows the user to specify the name of a place of interest using an address or zip code, which is then matched against relevant documents. A commercial enterprise devoted to geographical indexing of documents is the company MetaCarta [Met]. Some explanation of the workings of an experimental geographical search engine can be found in [Egn]. The software uses the US Bureau of the Census TIGER/Line digital mapping data to detect street addresses in a corpus of text and then converts them to geographical coordinates. These coordinates are indexed in a two-dimensional index along with a conventional keyword index of the corpus. A query processor is able to process queries that ask for documents which match certain keywords or contain addresses within a certain radius of a specified target address. The procedure for detection of geo-referencing appears to be used in conjunction with the manual registration of URLs in the GeoURL's location-to-url reverse directory database creation [Geo]. Other research which considers the specific problem of determining the geographical context of a web document is that of [DGS00]. In their work they used a gazetteer to detect to the presence of place names in a web document, before analysing their frequency. Another approach to developing location-specific referencing of web data is to associate IP addresses of domain names with telephone area codes as suggested by Buyukokkten et al [BCGM$^+$99]. In this approach the postal address of the web site or network administrators is used to derive a zip code that can be mapped to geographical coordinates. The Stanford Research Institute (SRI) has proposed a top level domain that is based on geographical referencing. In this system, the domain name refers to a strict hierarchy of quadrilateral cells defined by latitude and longitude. Existing domain names would be able to register themselves with a *.geo* domain server which would store, for a set of cells, all registered web sites that relate to each of the given cells. An experimental system for geographical navigation of the web has been described by McCurley [McC01]. A variety of techniques is proposed for extraction of the geographical context of a web page, on the basis of the occurrence of text addresses and post codes, place names and telephone numbers. This information is then transformed to one of a limited set of point-referenced map locations. Geographic search is initiated by the user asking to find web sites that refer to places in the vicinity of a currently displayed web site.

Global Atlas search engine [BOL00] indexes maps, images and HTML documents on the Web. The indexes are maintained for the available information according to their "geoprint" in addition to the traditional keywords and categories used by most search-engines. Queries are expressed as rectangles drawn on a map together with the traditional keyword filters. The registration of document footprints into an Oracle based spatial database is done with the help of gazetteers such as the Getty Thesaurus of Geographic Names [Get].

3 Architecture of the SPIRIT Search Engine

The SPIRIT search engine consists of the following components: user interface; geographical and domain-specific ontologies; web document collection; core search engine; textual and spatial indexes of document collection; relevance ranking and metadata extraction as shown in figure 1. We now provide a summary of the functionality associated with each of these components and the interactions between the components required to support the functionality. The summary starts with the user interface to introduce the system functionality from the user's perspective.

3.1 User Interface

The user interface allows the user to specify a subject of interest and a geographical location. The initial provisions for specifying a query are a structured text interface, a free text interface and a map. It is also planned that facilities for query by sketch will be introduced. The structured text interface allows the user to specify the subject of the query, a place name and a spatial relationship to the place name. The term or terms that form the subject of the query are treated in our initial prototype as non-spatial, but they could include types of places such as "hotels", or "cities."

The spatial relationships that are to be supported initially are *inside*, *outside* and *near* (distinguishing between whether the query is to include the named place or not) as well as cardinal direction and proximity relationships, namely *within a specified distance.*

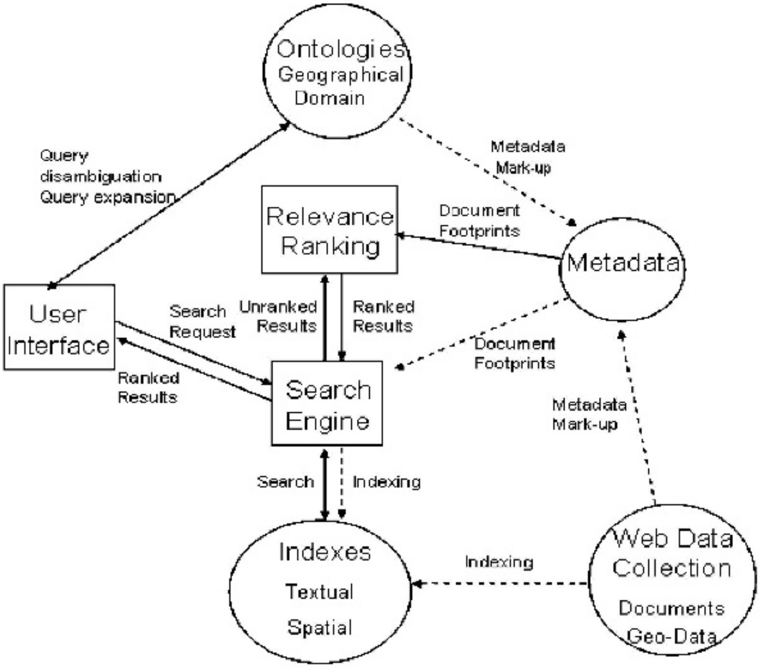

Fig. 1. SPIRIT search engine architecture.

Once the user enters a query, it is tested for ambiguity and alternative options are presented for purposes of disambiguation before the query is submitted to the search engine. The combination of the place name and spatial relationship is then used to determine a geometric query footprint, which consists of a polygon representing the interpretation of the spatial relationship applied to the place name. It might, for example, consist of the region of space inside a city or within some specified distance of a city. The shape of the footprint is echoed on the user interface map. The processes of disambiguation and query footprint generation require access to the geographical ontology. Once the user confirms the system's interpretation of the query, it is submitted to the core search engine and a relevance-ranked list of documents is returned to the user. These results will be presented both as a text list and as symbols on a map. A number of graphical visualizations are being investigated.

The present state of research in the field of spatially-aware search engines leaves many issues unresolved, concerning for example, effective user interfaces for geographical information retrieval, the most appropriate information to maintain in a gazetteer or a geo-ontology, reliable methods for determining the geographical scope of web documents, the way in which geographical relevance should be combined with thematic relevance to obtain a good ranking of retrieved documents, and the relative merits of different techniques for indexing documents with regard to both space and textual content. This paper provides one of the first published accounts of the overall architecture of a spatially-aware search engine that serves as a testbed for specific research to address these issues. It elucidates the key role of a geographical ontology in supporting geographical information retrieval functionality, and presents experimental results on a novel technique for integrating spatial and textual indexing for efficient access to web documents.

3.2 Ontologies

The primary ontology component is a geographical ontology that provides a model of the terminology and structure of geographic space. The geo-ontology plays a key role in the interpretation of user queries; the formulation of system queries, generation of spatial indexes, relevance ranking and metadata extraction. The design and implementation of the geo-ontology is described in more detail in Section 4.

In addition to the geographical ontology, the SPIRIT prototype also maintains a domain-specific ontology, which is focused on tourism. This will enable query expansion with respect to subject query terms for this domain. Thus if the user employs the term “accommodation” it will be possible to expand this to include terms such as “hotel,” “guest house” etc. It will also introduce the possibility of disambiguation with respect to the subject query terms (e.g., does “surfing” refer to wind surfing or surf boarding etc).

3.3 Web Document Collection

Initial experiments with the SPIRIT search engine employ a 1 terabyte test collection of web documents (comprising 94 million web pages). Documents in this collection are structured to facilitate indexing. In order to support spatial indexing of the collection, each referenced document that contains place names is associated with one or more document footprints that are derived from the geographical ontology entries for the respective names. The SPIRIT prototype adapts an experimental text search engine GLASS [GLA] for purposes of building, maintaining and accessing the document collection and indexes. The major modification to the existing search engine functionality concerns the introduction of spatial indexing of the indexes of web documents and facilities to search for geographical context within web documents.

3.4 Indexes

The SPIRIT search engine supports both pure text indexing and spatio-textual indexing. The text index employs a conventional inverted file structure whereby each term in the index is associated with a list of the documents that include the term. Use of this index alone to process both spatial and non-spatial terms of a query is analogous to conventional search engine functionality, and will depend for its geographical effectiveness upon exact match between query place names and place names in web documents. The results of such a search can be improved in principle by expanding the query terms using the ontology, as indicated above, to refer to synonyms and neighbouring places. Note that in some cases this may result in massive proliferation of query terms, as would be the case if the name of a large region were expanded to include all contained places.

Spatio-textual indexing combines text indexing with spatial indexing of documents with respect to their document footprint. Use of a geometric query footprint to access a spatial index of documents serves the same purpose as geographical term expansion, with a term index, in that all documents relating to the region of the query should be retrieved. Use of a query footprint to access the index avoids the potentially very heavy and possibly impracticable overhead of employing high numbers of query terms resulting from term expansion. Spatial indexing will also facilitate distance-based relevance ranking, which depends upon analysis of the geometric relationships between query and document footprints. Details of spatio-textual indexing are provided in Section 5.

3.5 Core Search Engine

This component is responsible for accessing the web document collection and its text and spatio-textual indexes. It is based on a simple text retrieval system that has been enhanced to support spatial access to web documents. The component receives a query from the user interface and processes it against the collection using the textual and spatio-textual indexes as required. For experimental purposes queries may be processed either entirely by means of the text index, or using the spatio-textual index. The initial results of the query are passed to the relevance ranking component before being returned to the user interface.

3.6 Relevance Ranking

The relevance ranking component takes results retrieved from the search engine database and relevance ranks them with respect to the non-spatial and spatial elements of the query. Text relevance ranking is based on the BM25 algorithm [RWB$^+$95], while spatial relevance is based, in the initial prototype, on measures of distance between the query footprint and the document footprint and on angular differences from cardinal directions in the case of directionally qualified queries. There are various techniques for combining textual and spatial relevance scores to produce an integrated score. In addition to combining independent text and spatial scores it is also possible to take account of the proximity in the document between the query text terms and the query spatial terms. This will be addressed in future versions of the SPIRIT prototype. It is also intended to introduce relevance ranking measures that take account of the parent geographical regions of the query footprint and the document footprint using methods such those documented in [CAT01].

3.7 Metadata Extraction

Effective spatial indexing of web documents depends upon the development of reliable techniques to identify the presence of place names in web documents and to determine their likely importance with regard to the subject matter of the document. This is the subject of ongoing research which includes the use of machine learning methods. For an example of previously published techniques see [DGS00,Li03]. Once significant place names have been detected in a document the geographical ontology can be used to provide footprints that contribute to the geographical metadata associated with the document. In the SPIRIT project, in addition to developing techniques for extraction of geographical context from web documents, work is also being pursued on the detection of features within geodatasets, which may then be used to enrich the geographical ontology [HKS03].

4 A Geographical Ontology for Information Retrieval

When interacting with the user, the geo-ontology is used to recognise the presence of place names in a query and then to perform disambiguation. Once the user's query is formulated as a $< term, spatialrelationship, place >$ expression, the ontology can be used to generate a polygonal geometric query footprint covering the spatial extent of the query region, based on the interpretation of the spatial relationship with the place. This query footprint is then used to access the spatial index of web documents. The geo-ontology could also be used to "expand" the user's query terms to include alternative names for the same place as well as the names of geographically associated places that may be inside, nearby or contain the specified place. The relevance ranking component accesses the geographical ontology to retrieve geometric footprints of places that are being compared with the query footprint, as well as with associated data providing the geographical context of a place, such as its containing and overlapping places. In the process of metadata extraction from web documents, the ontology is essential in identifying the presence of place names within text. There is also scope for enriching the ontology by including imprecise places found as a result of analysis of geo-datasets.

To support the functions above, actual and alternative place names,including multi-lingual versions need to be supported by the ontology. Geographical containment hierarchies and place types are required for query expansion and disambiguation. A geographic place is associated with possibly multiple geometric footprints. For example, detailed geometric footprints need to be used for accurate spatial indexing of documents. As indexing is a pre-processing operation, no impact on run time performance is expected. However, when generating a query footprint, access to detailed geometries can be expected to introduce processing overheads and hence there is a strong case for supporting generalised polygonal e.g. an MBR, or point-based geometries. The same reasoning applies to the use of a footprint for spatial relevance ranking of documents at query time. Consequently, the design of the geo-ontology supports multiple spatial representations of geographic places, including centre points, minimum bounding rectangles besides more faithful representations of geometries.

Several types of spatial relationships are stored and supported by the geo-ontology. Part-of relationships are used for maintaining containment hierarchies (e.g. based on different types of administrative hierarchies). Overlap and adjacency relationships are utilised by the relevance ranking component. In addition, overlap and containment relationships are used to derive similarity metrics between pairs of places. The later function is used when building or updating the ontology using different data sources. Hence, the geo-ontology supports part-of, contains, overlap and adjacency relationships between geographic places. The main components of the SPIRIT geo-ontology are shown in figure 2.

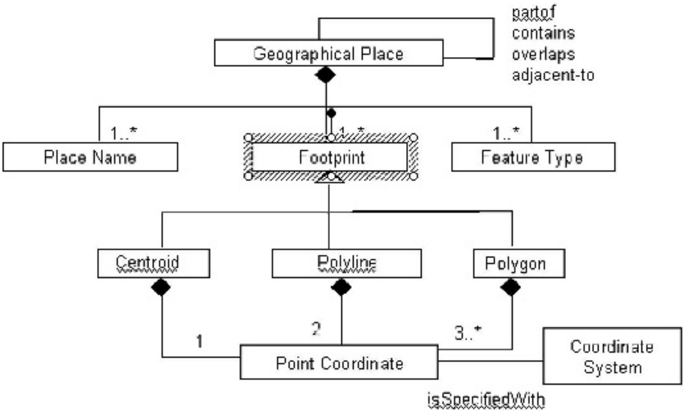

Fig. 2. Conceptual design of the geo-ontology.

The geographical ontology has been implemented using the Oracle Spatial database management system. The set of access functions developed are summarised as follows.

matchPlace: is an operation used by the user interface component to test whether a specific term in a query is a valid geographical place name.

getFeatureID: is an operation used by the geo-markup process to get the identifiers of places of specified terms.

getFootprint: is an operation to retrieve the geographical footprints for the specified place.

queryDisambiguation: is an operation to retrieve the broader geographical contexts for the name appearing in a query, using the partof relationship to derive the geographical hierarchies which are returned.

queryExpansion: is an operation that takes as arguments the disambiguated place name and the spatial relationship and derives the desired geographical search extent for the query (i.e. the query footprint).

Multiple data sources for different countries in Europe have been used to populate the geo-ontology. In particular, the EuroGeographics SABE data set [Sea] was used to extract the locations of towns and administrative boundaries and the Getty Thesaurus of Geographical Names [Get] is used to complement the ontology by providing other information such as alternative place names.

Employing multiple data sources in building the ontology is a complex task involving a pre-processing stage for checking the similarity of the data sets. Data sets may be different in many respects including the accuracy of representation, the type of spatial representation of the geographic objects as well as in the semantic classification used. Research in this area is still ongoing and will be reported elsewhere.

5 Spatio-textual Indexing

In this section an approach to integrating spatial indexing with textual indexing by means of spatio-textual keys is described. Preliminary experimental results with synthetic data compare the results of the spatio-textual indexing method with pure textual (PT) indexing. A spatial index of web documents can be created in a similar way to a spatial index of a geographic data set. Each document is allocated one or more geometric footprints, typically in the form of polygons. The footprints can then be referenced by the cells of spatial indexing methods such as a regular grid, a quadtree, or an R-tree. In the current implementation of the SPIRIT prototype, a regular grid scheme is employed. In this section, the derivation of spatio-textual keys is explained for an example document space as shown in figure 3. A collection of 16 documents, $D = \{D_1, D_2, \cdots, D_{16}\}$, is distributed over a document space R divided into 4 cells. The footprints of the documents (approximated by rectangular bounding boxes) are shown. Let SR be the document space associated with the entire set D, and the respective division for cells R_1, R_2, R_3 and R_4 be $SR1$, $SR2$, $SR3$, $SR4$.

$SR = \{D_1, D_2, \cdots, D_{16}\}$
$SR1 = \{D_1, D_7, D_{12}, D_{15}\}$
$SR2 = \{D_{15}, D_{10}, D_{11}, D_3, D_{13}\}$
$SR3 = \{D_2, D_5, D_{14}, D_{12}, D_{15}\}$
$SR4 = \{D_{15}, D_{14}, D_9, D_6, D_{11}, D_{16}, D_4, D_8\}$

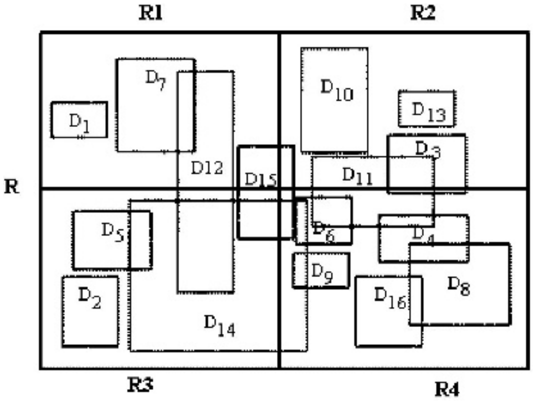

Fig. 3. An example of a document's search space.

A complete text index for SR corresponds to the index type PT. In such an index each term is associated with a list of the documents (inverted document list) in which it occurs. An example of such a list for a text term "spirit" is as follows:

spirit	$\{D_1, D_2, D_3, D_7, D_8, D_9, D_{11}, D_{13}\}$

Using the cell divisions shown in figure 3, this inverted list can be re-arranged in the form of a spatially-referenced list of documents which records for each cell those documents that contain the term:

spirit	$\{R1(D_1, D_7); R2(D_3, D_{11}, D_{13}); R3(D_2); R4(D_8, D_9, D_{11})\}$

An index with this structure can be exploited by first searching for a textual term and then using the associated spatial index of documents to filter out those meeting the spatial constraints.

A greater degree of integration of text and space, which effectively reduces the dimensionality of the problem, can be obtained by creating spatio-textual keys which join a text term to its corresponding spatial cells. Here the numeric cell identifiers ($R1$, $R2$ etc in the example) are concatenated with the text. In the example, the resulting keys would then be associated with their document lists as follows:

spiritR1	D_1, D_7
spiritR2	D_3, D_{11}, D_{13}
spiritR3	D_2
spiritR4	D_8, D_9, D_{11}

Searching this index requires text query terms to be concatenated at run time with the identifiers of spatial cells intersecting the query footprint, prior to matching the transformed query terms with the spatio-textual index terms. This indexing strategy shall be denoted spatio-textual (SP). Textual terms are used as a prefix for making the spatio-textual key. Identical results should be expected if cell identifiers were used as prefix, as the resulting spatio-textual list is of the same size and is constructed in exactly the same manner. It is possible for the spatio-textual index size to be bigger than the pure textual index, due to the introduction of multiple lists of documents for each term. The individual document lists may however be much smaller than for the PT index, provided that the terms referred to within documents in individual cells are a relatively small subset of the total number of terms in the document database. The following experiments were designed to predict the performance of the indexing strategies.

5.1 Experiments

In the absence of a very large collection of spatially indexed documents, a synthetic collection is used in which documents are assigned random footprints.

Queries are also generated synthetically. A total of six synthetic data sets were used comprising 1000 to 10,000 documents by employing techniques described in the public domain "mg" source code [MG.]. This text data generator allows the generation of data and queries with realistic properties as per their storage requirements. The document search space is divided into 1 to 60 cells to study the effect of cell size on search performance. As query time for textual queries up to 100 terms is practically undetectable, a query set comprising 1000 queries is used. The source code is written in C++ using MFC and STL in the Windows environment. A Pentium 4 PC is used (2 GHz processor and 256 MB of RAM). Standard MFC/STL arrays are used to store all indexes and table lookup is implemented with a standard binary search algorithm. A discussion of the results of the experiments are presented below.

5.2 Discussion of Results

The batch of 1000 queries was subjected to a total of 13 indexes for each of the six data sets. Table 1 lists the different index types.

Table 1. Different index types used in the experiments.

Index Type	Cells	Rows	Columns	Cell Size (%)
PT				
PT + S				
SP_R1_C1	1	1	1	100.0
SP_R2_C1	2	2	1	50.0
SP_R2_C2	4	2	2	25.0
SP_R5_C2	10	5	2	10.0
SP_R4_C5	20	4	5	5.0
SP_R5_C5	25	5	5	4.0
SP_R6_C5	30	6	5	3.0
SP_R7_C5	35	7	5	2.8
SP_R5_C8	40	5	8	2.5
SP_R5_C10	50	5	10	2.0
SP_R10_C6	60	10	6	1.6

In table 1, *PT* is a pure textual index comprising text terms only. All documents matching query terms are returned. The $PT + S$ index is the same as the *PT* index, but a spatial post-processing phase is employed to ensure that the only query results returned are those that fall within the query's footprint. This index is used to give an indication of query times and document hits when spatial relevance of documents is decided during query time. $SP\_R * \_C*$ are spatio-textual indexes with one or more cells.

Table 1 and figures 4 and 5 depict characteristics of the different index types. *SP* terms (Max) is the maximum number of *SP* terms obtained as a product of the total number of indexed terms (T) and the number of spatial cells (S).

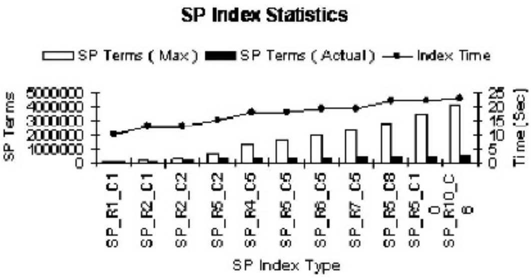

Fig. 4. SP index statistics.

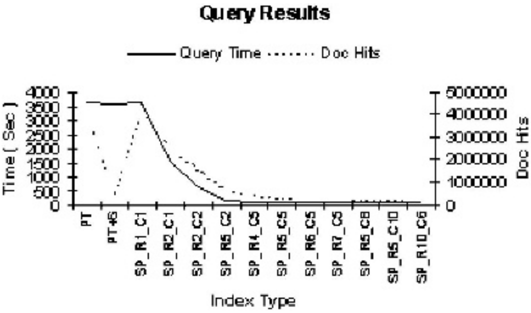

Fig. 5. Query results.

On comparing the theoretically expected maximum and actual *SP* terms, for each of the data sets, the actual number of text terms in the *SP* indexes is found to be much less than the expected maximum. Theoretically the number of *SP* terms should increase when the *SP* index has a higher number of cells. In practise this number is very low as can be seen in the graphs. This is a very positive observation in terms of memory considerations for spatio-textual indexes.

The *SP* indexes are constructed by post-processing the *PT* indexes and as such their construction times is much smaller in comparison. Figures 4 and 5 correspond to a collection size of 10,000 documents but a similar trend was also observed for other data sets.

As shown in figure 5, the query time for *PT* and $PT + S$ indexes is far higher than most SP indexes as the search space is the textual index of the entire collection. The only exception is the *SP* index SP_R1_C1, where there is only one spatial cell and the search space is identical to the *PT* and $PT + S$ indexes i.e. the entire collection.

The query time for the $PT + S$ index is slightly less than the *PT* index in all cases. Initially this may not sound logical as the $PT+S$ index performs additional calculations for filtering out the documents that do not intersect the query's footprint. However, observing the number of documents returned for both those index types it can be seen that the document hits after spatial filtering by the

$PT + S$ index are far lower than the PT index. Thus the lower query time is possibly due to fewer I/O operations used in writing the results.

The query times for the SP indexes initially begin to fall with increasing resolution of the grid but then decrease gradually. The document hits keep falling with increasing grid resolution. This suggest that excessive grid refinement can cause information loss. Comparing both time and document hits for the SP indexes with the $PT+S$ index, it appears that the best configuration of the grid is for SP_R4_C5 i.e. when the cell size is 5%. This corresponds to greater than 30 times improvement in query time.

6 Conclusions

This paper has summarised the architecture of a spatially-aware search engine and focused on the design of its component geo-ontology and on the integration of textual and spatial indexing of web documents. The prototype is currently under development and demonstrates the viability of the overall design. The geo-ontology plays a key role in providing support for query disambiguation, query expansion via the generation of geometric query footprints, relevance ranking to compare a query footprint to a document footprint, and the extraction of metadata to record the geographical context of web documents and geo-datasets. The introduction of spatial indexing, through the use of spatio-textual keys that concatenate text and spatial cell identifiers, reduced query times in excess of 30 times for large data sets for some grid resolutions. A study of the effect of the size of a spatial cell was conducted in the context of the regular grid indexing scheme and the best spatial search times were achieved when the cell size was 5% of the total area of the data sets employed.

Further work relating to the role of geo-ontologies and spatio-textual indexing will address issues that include integration of multiple data sources for construction of multi-national geo-ontologies, multi-scale representation of place footprints, representation of imprecise named places, alternative forms of spatio-textual key and experiments with large collections of geo-referenced documents. Research in parallel with that presented here focuses on issues concerned with user interface design, geo-referencing of text in web documents (to establish document footprints), geographical relevance ranking methods and metadata extraction from and enhancement of geo-datasets for purposes of web search.

Acknowledgments

This research is supported by the European Commission Framework V project SPIRIT: Spatially-Aware Information Retrieval on the Internet, funded under the Semantic Web Technologies action line.

References

[Ale] Alexandria Digital Library Project. http://www.alexandria.ucsb.edu/.

[BCGM$^+$99] O. Buyukokkten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar. Exploiting Geographical Location Information of Web Pages. In *Proceedings of Workshop on Web Databases (WebDB'99) held in conjunction with ACM SIGMOD'99*, pages 91–96. ACM Press, 1999.

[BOL00] S. Bressan, B.C. Ooi, and F. Lee. Global Atlas: Calibrating and Indexing Documents from the Internet in the Cartographic Paradigm. In *Proceedings of the 1st International Conference on Web Information Systems Engineering*, volume 1, pages 117–124, 2000.

[CAT01] Jones C.B., H. Alani, and D. Tudhope. Geographical information retrieval with ontologies of place. In *Spatial Information Theory Foundations of Geographic Information Science, COSIT 2001*, volume LNCS 2205, pages 323–335. Springer Verlag, 2001.

[DGS00] J. Ding, L. Gravano, and N. Shivakumar. Computing Geographical Scopes of Web Resources. In *Proceedings of the 26th Very-Large Database (VLDB) Conference*, pages 546–556. Morgan Kaufmann, 2000.

[Egn] Egnor, D. http://www.google.com/programming-contest/winner.html.

[Geo] GeoURL ICBM Address Server. http://geourl.org/.

[Get] Getty Thesaurus of Geographic Names. http://www.getty.edu/research/conducting_research/vocabularies/tgn/index.html.

[GLA] GLASS: Online Documentation. http://dis.shef.ac.uk/mark/glass/.

[Goo] Google Local. http://local.google.com/lochp.

[HKS03] F. Heinzle, M. Kopczynski, and M. Sester. Spatial Data Interpretation for the Intelligent Access to Spatial Information in the Internet. In *Proceedings of 21st International Cartographic Conference*, 2003.

[Li03] H. Li. Infoxtract location normalization: a hybrid approach to geographic references in information extraction. In *Proc. of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 39–44, 2003.

[Map] Mapblast. http://www.mapblast.com.

[McC01] K.S. McCurley. Geospatial Mapping and Navigation of the Web. In *Proceedings of Tenth International World Wide Web Conference*, pages 221–229. ACM Press, 2001.

[Met] Metacarta. http://www.metacarta.com.

[MG.] MG. Information Retrieval System. http://www.cs.mu.oz.au/mg/.

[Mir] Mirago: Mirago the UK Search Engine. http://www.mirago.co.uk/.

[Nor] Northern Light. http://www.northernlight.com/index.html.

[RWB$^+$95] S.E. Robertson, S. Walker, M.M. Beaulieu, M. Gatford, and A. Payne. Okapi at trec-4. In *Proc. of the 4th Text REtrieval Conference (TREC-4)*, pages 73–96, 1995.

[Sea] Seamless Administrative Boundaries of Europe (SABE) dataset. http://www.eurogeographics.org/eng/04-sabe.asp.

[Vic] Vicinity.com. http://home.vicinity.com/us/mappoint.htm.