Spatio-Textual Indexing for Geographical Search on the Web

Subodh Vaid[1], Christopher B. Jones[1],
Hideo Joho[2] and Mark Sanderson[2]

[1]School of Computer Science, Cardiff University, UK
email: {c.b.jones, subodh.vaid},@cs.cf.ac.uk
[2]Department of Information Studies
University of Sheffield, UK
email:{h.joho,m.sanderson}@sheffield.ac.uk

Abstract. Many web documents refer to specific geographic localities and many people include geographic context in queries to web search engines. Standard web search engines treat the geographical terms in the same way as other terms. This can result in failure to find relevant documents that refer to the place of interest using alternative related names, such as those of included or nearby places. This can be overcome by associating text indexing with spatial indexing methods that exploit geo-tagging procedures to categorise documents with respect to geographic space. We describe three methods for spatio-textual indexing based on multiple spatially indexed text indexes, attaching spatial indexes to the document occurrences of a text index, and merging text index access results with results of access to a spatial index of documents. These schemes are compared experimentally with a conventional text index search engine, using a collection of geo-tagged web documents, and are shown to be able to compete in speed and storage performance with pure text indexing.

1 Introduction

The main focus of developments in spatial database design has been in support of the maintenance of highly structured map-based geometric data and their attributes. The World Wide Web introduces a challenge to spatial databases in that it consists of a vast repository of largely unstructured information that is dominantly in the form of text documents. A large amount of information on the web is geographically specific, in the sense that it refers to particular geographical locations, but the geographic references are as a rule embedded within the textual content of the documents, in the form of place names, addresses, postcodes and the associated geographical terminology of spatial relationships. Users of the web often submit geographical enquiries requesting information about, for example, services relating to retailing, tourist attractions, accommodation, sport, entertainment, transport, public services and cultural heritage. In a study of a log of the Excite search engine, it was found that about one fifth of all queries were geographical, as determined by the presence of a

geographical term such as a place name, a post code, a type of place or a directional qualifier such as north [18].

When a user submits a geographically-specific web search they usually use a place name to provide the geographic reference. This name will then be treated the same as the other search terms and documents containing the query terms will be retrieved. For purposes of geographic search, this approach has major limitations in that it will ignore potentially relevant documents that refer to the place of interest but do not include the specified place name. Thus relevant documents could refer to places that are inside or near the specified place or they could use an alternative version of the specified place name. It is also the case that there are many places in different locations sharing the same name, resulting in the return of irrelevant documents. Another problem with using place names for geographic search with conventional search engines is that place names are commonly used in the names of organisations, people and buildings, resulting in the retrieval of documents that may have no geographical relevance despite the inclusion of the place name. In theory, the limitations above that result in missing relevant documents can be overcome by creating an expanded list of query terms. The expanded list could include alternative names and the names of places inside and nearby the target geographical location. In practice this would lead to the possibility of intractable query expressions containing many thousands of geographical terms. This would occur for target places that were spatially extensive to the extent that they contained many other named places. The approach would also inevitably result in the return of irrelevant documents that used the target place names to refer to the names of organizations, people or other phenomena for which the name does not provide geographical context.

There is a need therefore to develop geographically-aware web search technology that can index and retrieve effectively documents according to their geographical context. Indexing documents according to their geographical context would not only overcome the problems referred to above. It would also facilitate intelligent interpretation of spatial relationships that the user may employ to qualify the query place names. This includes terms such as *near*, *north of*, and *within 5 kilometers of*. Several working and experimental systems for geographical web search have appeared in recent years (some examples of which are reviewed below) but there is much work to be done to create effective systems. There are several important aspects of geographically-aware search that introduce challenges in their own right. Categorisation of web documents according their geographical content (geo-tagging) requires geo-parsing and geocoding procedures to detect and interpret geographical terminology in web documents and to "ground" (geocode) the resulting references with coordinates. This process of document categorisation requires a source of place name knowledge in the form of a gazetteer or geographical ontology that maintains information about place names in association with, for example, alternative names, geometric footprints that give coordinates for places, place types and the hierarchical structure of geographic space. Once documents have been categorised geographically they must be indexed with respect both to the textual content and to geographic space. Retrieval of documents must be accompanied by relevance ranking that needs to take account both of the geographical context and of the non-spatial concept terms that the user has employed in a query. Effective geographical search also requires a user interface that can help the user to disambiguate place names that refer to multiple

places and to assist in formulating geographically-specific queries and reporting the results.

In this paper we focus on the issue of combining text indexing with spatial indexing. We present two spatio-textual indexing schemes which may be regarded as spatial-primary and text-primary respectively and compare them with each other and with using a pure text index in conjunction with a separate spatial index of documents and with a pure text index by itself. Experiments are conducted in the context of the SPIRIT prototype spatially-aware web search engine [20], using a collection of actual web documents. The performance of the various schemes are compared with respect to index costs and to query times and the numbers of documents retrieved. In the remainder of the paper we summarise briefly previous work on geographical web search, before providing an overview of the architecture of the SPIRIT system. In Section 4 we describe the indexing schemes that have been implemented for these experiments, before reporting on results of applying them using several types of query in Section 5. The paper concludes in Section 6 highlighting the relative merits of the implemented techniques and indicating future research directions.

2 Related work

A geosearch tool from the Vicinity company was implemented in association with the Northern Light search engine [14], but no longer operates. It provided the facility to search for web documents on a specified topic relating to an address in the USA or Canada, allowing the user to specify a radius of search, The documents may have been spatially indexed, but the techniques employed were not openly published. Google introduced a geographical search facility in the Google Local version of their search engine [6]. Again, no explicit details are published on the spatial indexing methods. The search engine is associated with a Yellow Pages like business directory, allowing users to search for businesses in a geographic area using a wider range of keywords to search for the businesses than are stored with a typical Yellow Pages directory. In Europe the Mirago web site [13] provides a geographically specific search facility that allows the user to perform web searches based on administrative regions, which are also displayed on a map on the user interface. Sagara and Kitsuregawa [17] have described briefly a system for retrieving and scoring geographically specific documents from the web with a prototype spatial search engine. In a manner apparently similar to Google Local Search, they used Yellow Pages to generate key words to find documents on the web relating to listed businesses. These were then scored, according to measures of popularity and reliability, and indexed within the web search engine, but the indexing methods were not described. An experimental system for geographical navigation of the web has been described by McCurley [12]. 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. An early example of developing

methods to determine the geographical scope of web pages was described by Buyukokkten [3]. This involved associating IP addresses with telephone area codes of the associated network administrators, and hence, via zip code databases, to place names and geographical coordinates. The approach facilitates the analysis of the geographical distribution of web sites. For purposes of information retrieval it appears to require that the content of a web page is related to the place where the web page was created, which is not always be the case. Ding et al [5] attempted to determine the geographic scope of pages using a gazetteer to recognise the presence of place names which were then analysed with respect to their frequency of occurrence. They also considered the geography of the sources that linked to the web document. Silva, Martins, Chaves and Cardosa [19] described methods for determining of the scope of web documents in the Portugese tumba! web search engine. After transforming web documents to a structured XML/RDF format they were progressively augmented with geographical descriptors through a sequence of lexical analysis, geographical entity recognition and semantic and web inference procedures.

Recent work on establishing the geographic scope of web pages has been presented by Amitay et al [1]. They identify the presence of candidate place names using a gazetteer, before assigning confidence levels to the interpretation of the name based on associated evidence. For example, two ambiguous places used in the same document are likely to refer to the same parent region, and an ambiguous name when used multiple times is likely to refer to the same place each time. Following disambiguation, the geographic foci of a web page are determined based on analysis of the frequencies of occurrence of place names in association with knowledge of the geographic hierarchies. A technique for indexing web documents geographically using spatio-textual keys was presented briefly in [7] and evaluated using synthetic data. In the context of a synthetic web document collection, the approach was shown to be beneficial, but no evidence was provided for its accuracy when applied to real data. It may be noted that a large proportion of recent published research relating to geographic web search has been concerned with the problems of geotagging rather than issues of indexing the resulting geotagged documents. See for example [10] [15].

3 Overview of SPIRIT search engine

The spatio-textual indexing methods described in this paper were implemented in the experimental SPIRIT search engine [9] [7]. Here we describe briefly the overall architecture of the SPIRIT search engine in order to place the indexing methods in context. The main components are the user interface, document analysis and metadata extraction, core search engine, indexes, the geographical ontology, and relevance ranking. The user interface allows users to specify a concept, a geographical place name and a spatial relationship to the named place. Spatial relationships may be proximal (distance), topological or directional. Examples of types these types of queries are illustrated in Table 1.

Query Type	Example
1. Distance	1. schools *within 10 km* of Zurich city centre
	2. hotels *near* Cardiff University
2. Topological	1. hospitals *in* London
3. Directional	1. holiday resorts *north of* Milan

Table 1. Query types for a SPIRIT query

SPIRIT employs disambiguation functionality, to allow the user to select the appropriate instance of a place name that has multiple occurrences, and presents the search results as a list of URLs and on an interactive map linked to the retrieved document list. The geographical ontology stores knowledge of instances of place names with alternative names, place types, qualitative spatial relationships to other places and one or more geometric footprints giving an approximate spatial extent for the place [8]. Place footprints may be in the form of a representative point (centroid), a minimum bounding box, a polygon or a line. The user interface component uses the geographical ontology for disambiguation of the part of a user's query that specifies place. This results in a query footprint F^Q that represents a geometric interpretation of the user query with respect to the spatial relationship to the named place. SPIRIT supports query footprints in the form of minimum bounding rectangles and convex hulls. For many geographical queries, notably those that employ the "near" relationship, the user can be expected to be interested in documents that relate to locations in the vicinity of the specified geographical location as well as those that match exactly with the named place. To accommodate this, the query footprint is expanded beyond the boundary of the footprint of the specified geographical location. The resulting query footprint, along with the other textual query terms specifying the concept of the query $T = \{t_1, t_2, \ldots t_m\}$, is submitted to the search engine to determine a match with the indexed documents. In general a user query Q is transformed to the form $Q = T \cup F^Q$ prior to submission to the search engine.

The document analysis and metadata extraction component is used to build a database of web documents that are indexed with regard to textual content and to geographic context. The geographic context is encoded in the form of a document footprint F^d derived from footprints of place names in the geographical ontology that have been detected as geographically significant. The individual footprints of a document footprint are equivalent to the place name footprints in the ontology and are used to perform spatial indexing of the document. Typically there will be several individual footprints in a $F^d = \{f_1, f_2, \ldots f_n\}$ and hence a document may be spatially indexed with respect to multiple locations. The core search engine finds those documents whose footprints intersect the query footprint. The individual documents returned d_i consist of those documents in the document collection D which contain all the non-spatial textual query terms $t_j \in T$ and which have footprints that intersect the query footprint. The set of documents returned is therefore

$$\{d_i \mid d_i \in D, (t_j \in d_i \ (\forall j \in 1..m)) \wedge (Q \cap f_k)_{,} \ f_k \in F^d i \ \ (k \in \ 1..n) \}$$

where $F^d i$ refers to the document footprint of document d_i.

Relevance ranking determines an overall ranking for a document by combining a score from text ranking, in the form of a BM25 score [16], with a score from spatial ranking. The spatial ranking can be performed in several different ways. It measures

the distance between the query footprint and the document footprint(s) primarily as a Euclidean distance but it is also possible to measure angular difference in order to process queries that employ a directional spatial relationship. The textual and spatial scores can be combined using distributed and non-distributed methods [11].

It should be noted that retrieval of the set of document ids whose footprints match the query footprint is not accompanied by any geometric filtering prior to submission to the relevance ranking component. If a spatial indexing method is used in which documents are referenced by spatial cells, all documents referenced by a cell that intersects the query footprint are passed to the relevance ranking component. This is justifiable in that documents that are outside the query footprint will be ranked lowest in the geographical dimension, and will be geographically adjacent to documents within the query footprint.

4 Spatial and textual indexing

Here we investigate hybrid indexing schemes that combine inverted files, that list the documents containing indexed document terms, with a spatial access method to maintain the geometric footprints of indexed documents. The spatial indexing methods employed here are all based on a fixed grid scheme. Clearly more sophisticated spatial access methods could be used but a fixed grid lends itself to relatively simple schemes that should be sufficient to demonstrate the relative merits of the approaches presented (note that fixed grids are used successfully in some commercial GIS).

Once a textual index for terms and a spatial index for document footprints are available then either of them can be used first to get a set of results that can be refined by using the other. Thus an important issue is to decide the order of the search on the index types i.e. Text followed by Spatial or Spatial followed by Text. Here we present and implement schemes based on both approaches and compare their performance experimentally with each other and with a pure text indexing scheme. The pure text indexing scheme PT treats geographical terms the same as other text terms and hence relies entirely on exact matching of query terms with document terms. Our first spatio-textual scheme ST uses a spatial index in a first stage and later searches text indexes created separately for each cell of space. Access to the second spatio-textual scheme TS starts with a term index and then exploits spatial indexes associated with each term of the term index. The third scheme T performs textual indexing and spatial indexing of documents independently, before combing the results.

4.1 Pure text indexing PT

In the pure text indexing scheme an inverted file scheme is used consisting of a lexicon file, each record of which contains fields for an item of text and a pointer (and associated offset data) to an entry in the “postings” file containing lists of occurrences of those documents from the document set D of size N that contain the text item. There will be L records in the lexicon, where L is the number of indexed terms, and L lists of document ids in the postings file. In a worst case scenario, all documents

contain all indexed terms so that the list of document occurrences for a term would be of length N, resulting in O($L.N$) storage. (Note that we are including some component factors of some linear complexity functions, such as in the latter expression, in order to help make distinctions between the various indexing schemes). In practice it is generally assumed that following Heap's Law [2] the size of the lexicon is O(N^{β}), where $0 < \beta < 1$ with typical values between 0.4 and 0.6, with the occurrences storage being O(N). Total storage may therefore be regarded as O(N).

Queries to this index contain all the terms in the user's query, consisting of m non-spatial textual query terms and n geographical query terms. Assuming that, having found a text term, the cost of a read into memory of the corresponding document list is proportional to K_a, the maximum number of documents referenced by a lexicon term, then, if the lexicon is managed with an access structure such as a binary tree or a B-tree, the access time for the PT index is O($(m+n)(\log L + K_a)$).

4.2 Spatial primary index ST

In this index, the space corresponding to the geographical coverage of the place names specified in documents is divided into a set of p regular grid cells $C = \{c_1, \ldots. c_p\}$ and for each cell an inverted text index is constructed. Each text index is structured in the same manner as the pure text index PT described above, but the document set S that it refers to consists of those documents d_j whose document footprint $F^D j$ intersects the corresponding spatial cell. Thus for a particular cell c_i the corresponding documents $S = \{ d_j \mid d_j \in D \wedge F^D j \cap c_i \}$. Those documents whose document footprints intersect more than one cell will be represented in multiple cell text indexes. The principle of the ST index is illustrated with respect to the set of documents whose footprints are represented as rectangles in Figure 1. Here a collection of 16 documents, $D=\{D_1, D_2, .., D_{16}\}$, is distributed over a document space R divided into 4 cells. Let S_R be the document space associated with the entire set D, where the respective subdivisions for cells *R1, R2, R3* and *R4* are $S_{R1} = \{D_1, D_7, D_{12}, D_{15}\}$, $S_{R2} = \{D_{15}, D_{10}, D_{11}, D_3, D_{13}\}$, $S_{R3} = \{D_2, D_5, D_{14}, D_{12}, D_{15}\}$, $S_{R4} = \{D_{15}, D_{14}, D_9, D_6, D_{11}, D_{16}, D_4, D_8\}$.

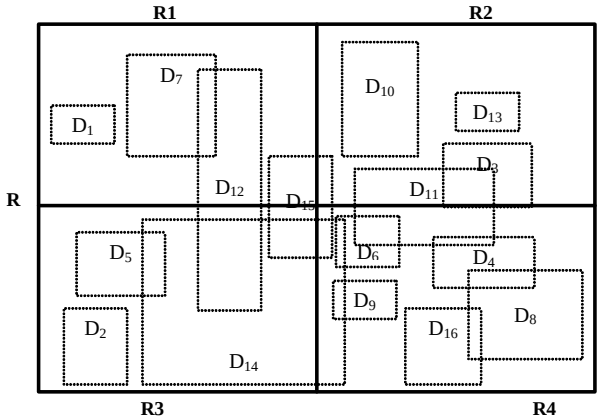

Figure 1. A spatial index of documents with rectangular footprints

In a worst case scenario the storage cost for this scheme would be p times that of the PT scheme, i.e. O($p.N$), corresponding to the event that all document footprints intersected all cells. In practice the process of categorising documents geographically, or geo-tagging, associates the majority of documents with a number of specific areas, reflecting the geographical focus of the documents. Consequently, the individual cell text indexes can be expected to be smaller than the PT index (this is investigated in the experimental results). Query times can be expected to depend on the number of cells r that are intersected by the query footprint and the sizes of the text indexes associated with those cells. Having determined which r cells are intersected by the query footprint, which can be computed relatively trivially in a regular grid spatial index, the subsequent query time would be $O(r((m\log(L_a) + K_b) + \log(p)))$, where m is the number of non-spatial query terms, L_a is the maximum number of terms indexed by the cell-specific indexes and K_b is the maximum number of documents referenced by a term in a cell-specific term index . Note that log(p) refers to the cost of locating the start of a cell-specific term index, and assumes a sorted list of location codes to identify the spatial cells. If the indexes are stored in separate files then access to the relevant index may be achieved in constant time.

4.3 Text primary spatio-textual index (TS)

In this index a pure text index structure is modified so that the list of documents in the postings file for each term is associated with a spatially-grouped set of documents that contain the term. The spatially organised documents take the form $[Cell_1[DocumentList_1]; Cell_2[DocumentList_2]….Cellp[DocumentList_p]]$, where $Cell_i$ are the cell identifiers of the regular grid cells and $DocumentList_i$ are the lists of documents whose footprints intersect the corresponding cell. For the example given in Figure 1, let us suppose that we have a list of documents associated with the index term “spirit” :

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

In the TS index the term “spirit” would be associated with a list of document occurrences grouped as:

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

In the worst case, the document footprints of all documents would intersect all cells. This storage may be characterized as O($p.N$). As indicated above in the context of the space primary scheme, in practice each document can be expected to be referenced by a subset of the cells, reflecting its geographical focus.

A query to this index consists of the m text terms and the query footprint. Having calculated which r spatial cells are intersected by the query footprint, m queries are required to the main index, corresponding to the individual textual query terms. For each such query, the r cells of the term-specific spatial index are accessed. If, for each term, there are maximum K_c documents referenced per cell then the access time is $O(m (\log(L) + r(\log(p) + K_c)))$.

4.4 Text index with spatial post-processing (T)

In this scheme we use a pure text index to find those documents D1 that contain the non-geographical query terms. Separately, a spatial index of documents (based on their footprints) is used to find those documents D2 whose document footprint intersects the query footprint. These two sets are then intersected to find those documents that both contain the non-spatial query terms and have a footprint that intersects the query footprint. The storage for this scheme is O(N) for the text index and O($p.K_d$) for the spatial index, where K_d is the maximum number of documents referenced by a spatial cell. It may be noted that the storage for ST might also be characterized in the same way, but there is a difference in practice in that the storage for ST is very much bigger, as for each cell a term index is stored, as opposed to the single list of document occurrences per cell in the T scheme.

The query time for accessing the text index is reduced relative to PT in that the geographical terms are not included, giving O(m(log$L$ + K_a)). The query time to access the spatial index of documents is O(r(log(p) + K_d)), where r is again the number of cells intersected by the query footprint. Thus each access to a spatial cell will be accompanied by a retrieval of the list of documents referenced by the cell. Having obtained two lists of documents they can be matched to find the common documents, a process that will be enhanced if the documents are stored in both lists in order of their document ids. In this case the match time would be directly proportional to the total numbers of documents.

5 Experimental comparison of indexing schemes

The performance of the four indexing schemes described above has been compared with regard to query time and to the numbers of documents that are returned. The spatio-textual schemes are compared with respect to differing cell size of the spatial index. The number of documents returned is of particular interest as this measures the size of the set passed to the relevance ranking procedure, which is itself a significant cost in the document retrieval process. We do not compare the quality of the results between the schemes from a user's point of view. This would require a geographical test collection of documents that had been ranked manually or semi-manually with regard to their relevance for particular queries. At present no such test collection is freely available though efforts are in place to create one [15]. In the present study all three spatio-textual schemes return exactly the same sets of documents for each spatial cell resolution that is studied. As explained earlier, the pure text scheme will inevitably be inferior with regard to a "recall" measure of quality in that, assuming no query term expansion, it will not find documents that are geographically relevant but which do not include the geographical query term employed in the query.

The document collection consists of 19,956 HTML documents relating to the United Kingdom taken from a terabyte-sized crawl of the Web conducted in 2001.A subset of 19,046 documents were allocated document footprints (geo-tagged) using

the GATE (General Architecture for TEXT Engineering) information extraction system [4]. ANNIE, the default Information Extraction system, is used to perform named entity recognition to detect the presence of place names. This uses gazetteer lists (e.g. common names of people and places) and context rules to disambiguate between named entities. These rules assist in distinguishing between place names that are used in a geographical context, and hence are of interest, and those that may be geographically spurious in that they refer for example to people's names and the names of organisations and buildings. The standard GATE gazetteer is enhanced here with the UK Ordnance Survey 1:50,000 gazetteer containing over 250,000 place names of topographic features and settlements, the SABE geo-dataset for the UK, from which more than 10,000 names and footprints were extracted, and the UK Ordnance Survey CodePoint dataset which lists more than a million UK postcodes.

Text indexing facilities are provided by an in-house research IR system called GLASS. All indexing schemes are file-based resulting in much longer access times than for a commercial system, in which most indexes would be maintained in main memory. As the purpose is to compare performance characteristics of spatio-textual indexing methods with pure text indexing, absolute timings are not of particular consequence. For each of the spatio-textual schemes, spatial cell resolutions range from a 2 X 2 subdivision of the geographical region covered by the geo-tagged documents to an 8 X 8 subdivision (and include a 1 X 1, i.e. single cell, subdivision for reference purposes). For each cell resolution we report statistics on the index sizes, the average numbers of documents referenced per cell and the average number of terms indexed within each cell. The purpose of these statistics is to demonstrate the way in which spatial indexing focuses search on geographically-specific documents.

5.1 Implemented indexing schemes

5.1.1 PT : pure text

The pure text indexing scheme employs the basic GLASS text indexing procedure that is exploited in the SPIRIT search engine. It follows the structure explained above and the file-based lexicon is accessed using a binary search on the sorted index terms. Query expressions include all geographical and non-geographical terms.

5.1.2 ST : Space-primary spatio-textual indexing

This scheme consists of a set of spatial cell-specific text indexes. Each such text index is implemented using the same indexing method as in PT, except that the documents indexed in an individual cell-specific index are those whose footprints intersect the cell. Following calculation of the cells intersected by the query, the files containing each of the relevant text indexes are accessed initially through the unix file system, with the file names being generated from the cell ids.

5.1.3 TS : Text-primary spatio-textual indexing

This indexing scheme is created by modifying the document occurrences lists in the GLASS index. For an individual indexed term, the occurrences list is segmented into cell-specific sub-lists. Each such sub-list contains the identities of documents whose footprint intersects the respective cell. The beginning of the occurrences file contains header data providing the offsets of the start and end of each cell-specific sub-list, supporting direct access reads to the relevant file sections.

5.1.4 T : Separate text and spatial indexes

In the T indexing scheme the pure text index component is identical in structure to that of PT, while the spatial index consists of a table containing records with the structure [cell_id, document_list]. The unix *grep* command is used to the access relevant parts of the file for a given cell id in order to read the respective sub-list into main memory. This may fall short of the theoretical logarithmic access referred to above in this context. The results from the text and spatial index, which are ordered by document id, are intersected using a unix shell script matching procedure.

5.2 Query schemes

Four query sets were employed , for each of which 100 queries were run. We now describe these query sets.

5.2.1 Query Set 1 : Random text terms and random place names (Random)

Non-geographical concept query terms were selected randomly from the terms in the lexicon and combined with a randomly selected geographical term selected from the SPIRIT list of geographical place names within the UK region. The number of non-geographical query terms was also chosen randomly from the range of 1 to 10.

5.2.2 Query Set 2 : Selected concept terms and random geography, largest 500 footprints (Top500FP)

The non-geographical query terms were selected randomly from 241 concepts (terms or phrases) obtained from the UpMyStreet.com web site, which provides a directory of geographically-specific information. The geographical terms were chosen randomly from the 500 SPIRIT UK place names with the largest footprints. These queries will tend towards larger geographic areas, using “realistic” concept terms.

5.2.3 Query Set 3 : Selected concept terms and random geography, smallest 500 footprints (Bottom500FP)

This query set adopts the same approach as Query Set 2 except that the geographical terms are now those in the SPIRIT geo-ontology with the 500 smallest footprints. In this case the geographical search is highly focused and would be expected to lie often within a single cell of the spatial indexes.

5.2.4 Query Set 4 : Selected concept terms and random geography from 5 largest footprints (Top5FP)

This query set takes concept terms as in Query Sets 2 and 3, but it may be regarded as an extreme version of Query set 2 in that the query footprints are derived randomly from the 5 SPIRIT UK place names with the largest footprints. It will tend therefore to maximise the numbers of spatial cells that need to be accessed to retrieve relevant documents.

5.3 Experimental results

The experimental results have been used to compare the schemes with regard to the size of the indexes, the time to construct the indexes and the query times for each of the four query sets. Results are presented with respect to the differing spatial index resolutions. We also show how the numbers of documents returned, the numbers of documents that intersect each spatial cell, and the numbers of terms indexed, change with cell size.

The sizes of the indexes for each of the schemes are compared in Figure 2. Here we can see that, for the ST and TS schemes, decreasing cell size, and hence increasing numbers of cells, has a significant negative impact on storage, as predicted in Section 4. For the highest resolution index with $p = 64$ cells, the latter schemes are in fact about 20 times bigger than the PT scheme. This factor demonstrates that there is a definite degree of geographical focus of the documents. This focus is illustrated in Figure 3 which plots the average numbers of documents and of terms per cell against grid resolution. Note that for 8 X 8 grid resolution there are about 3000 documents per cell on average, out of a total of nearly 20,000 documents. This reflects the fact that many documents are represented by multiple individual footprints, averaging 21, with a maximum value of 803 in these experiments. The T scheme shows very little degradation in index size with increasing grid resolution. This is because the spatial index of documents, used here with the PT index, occupies relatively little space compared with the term indexes. The total storage for the 8 X 8 resolution grid index is about 1Mb, whereas the total storage for PT is of the order of 100Mb (see figure 2). Note that that in this and subsequent figures “GLASS” in the legend refers to the PT scheme.

The indexing times for the schemes are presented in Figure 4. The ST scheme stands out as having poorer performance with increasing grid resolution. This scheme differs notably from the others in that it is necessary to build separate inverted text indexes for each spatial cell. TS in comparison is more integrated, with a single text index. It is the document occurrences file that is modified in TS relative to PT, with the additional cell-specific document occurrence “sub-lists”.

Table 2 summarises some statistics of the four query sets that were used to study query timings and numbers of documents retrieved for the different indexing schemes. It presents the minimum, maximum and average size of the query footprint

as a percentage of the total area of the indexed region. Note that for the highest resolution spatial indexing scheme, each grid cell would be about 0.016%

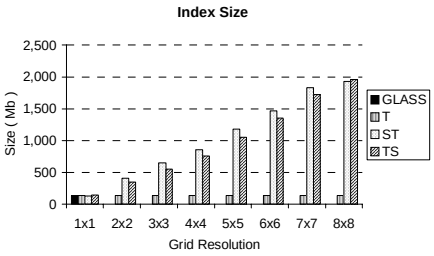

Figure 2. Index Size (GLASS refers to PT)

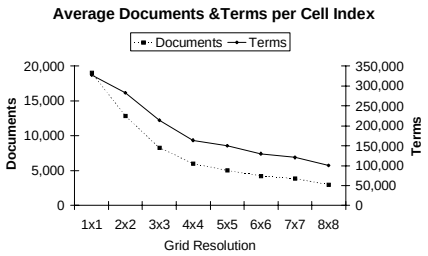

Figure 3. Average documents & Terms per cell

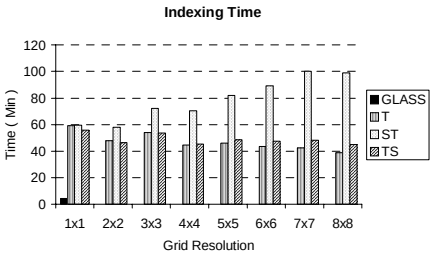

Figure 4. Indexing Times

Thus in the Top500FP query set the biggest query footprint is similar in size to the smallest cell that is used in the spatial indexing method. The third query set Bottom500FP has small query footprints relative to spatial index cell size, while the query footprints for the fourth query set Top5FP are extremely large, averaging 0.39 of the entire indexed region. The number of terms in the concept part of the queries is given and it corresponds to the value *m* in the theoretical discussion in Section 4. The information for numbers of terms in the place name corresponds to the value *n*, and as such it affects only the PT indexing scheme, since these terms are not submitted to the indexes directly in the other schemes (they are converted to a geometric query footprint). All queries use the "near" spatial relation and as such result in an increase in size of the footprint of the target place name in order to generate the query footprint.

Query Set		Query Footprint (% of Total Space)	Terms in Concept	Terms in Place Name
Random	*Min*	0.000395	1	1
	Max	0.069283	10	4
	Average	0.002951	5.94	1.48
Top 500 FP	*Min*	0.000399	1	1
	Max	0.017344	6	4
	Average	0.002185	2.65	1.61
Bottom 500 FP	*Min*	9.47E-08	1	1
	Max	1.67E-06	7	3
	Average	1.21E-06	2.87	1.55
Top 5 FP	*Min*	0.061869	1	1
	Max	1	6	2
	Average	0.391055	2.55	1.2

Table 2. Query set characteristics

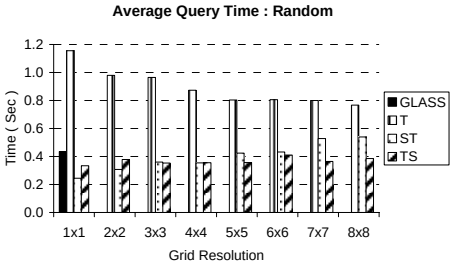

Figure 5. Average Query Time : Random

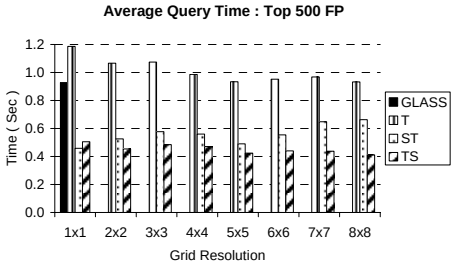

Figure 6. Average Query Time : Top 500 FP

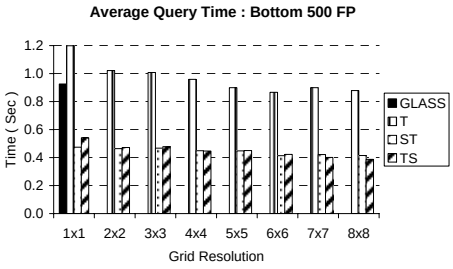

Figure 7. Average Query Time : Bottom 500 FP

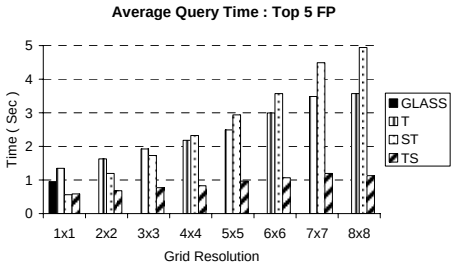

Figure 8. Average Query Time : Top 5 FP

Figures 5 to 8 illustrate the average query times for each query set respectively, based on 100 queries for each query set. In the first three query sets, the ST and TS schemes are similar or better than PT for all grid resolutions. The T scheme stands out as somewhat inferior to the other schemes, though in the case of the more realistic query sets (Top500FP and Bottom 500FP) it is usually no worse than double the other spatio-textual schemes and not much worse than PT. We regard this poorer

performance as a reflection of an inefficient document merging process (using unix shell scripts) that matches the results of the pure text index with the spatial index of document occurrences.

Figure 8 illustrates the results for the Top5FP query set, which employs very large query footprints. The results here clearly reflect the theoretical analysis whereby timings depend upon the numbers of spatial cells intersected by the query footprint. The average query footprint occupies 0.39 of the entire index and hence intersects a similar proportion of the spatial cells of the respective indexes. This impacts most upon schemes ST and T in which results must be obtained from each intersected cell, prior to merging of the result sets. In both ST and T the merging is performed outside of the main index access programs, using unix scripts. The TS scheme works comparatively well with this query set as all data processing is performed within the shared memory of the modified version of GLASS. In this respect it is the most well integrated spatio-textual scheme. The absolute query times for all schemes here were slow (about a second per query), but this is due to the use of disk-based as opposed to main memory storage methods and the fact that the text indexing methods are not optimised in several respects.

Figures 9 to 12 illustrate the numbers of documents returned for each of the query sets. In the Random query set only about 2 documents are being returned per query, due to the unrealistic random combinations of concept query terms, and no clear pattern emerges. The other three schemes, notably Top500FP and Bottom 500FP, demonstrate a clear trend of reducing numbers of documents returned as grid cell resolution increases. The reason for the decrease in numbers of documents returned is that, as indicated previously, there is no filtering at this stage of the retrieved data against the query footprint. All data in spatial cells that intersect the query footprint are returned. As cell size decreases so there will be a decrease in the numbers of documents that fall outside the query footprint but which lie inside the intersected index cells. The fewer documents that are outside the query footprint the less work is required of the relevance ranking component. In the results here for the Top500FP and Bottom500FP, i.e. the most realistic query sets, the highest resolution spatial indexes result in returning about 50% of those documents returned using a single cell.

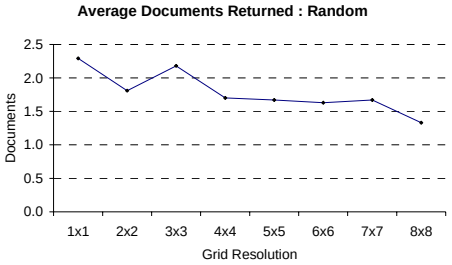

Figure 9. Average Documents Returned : Random

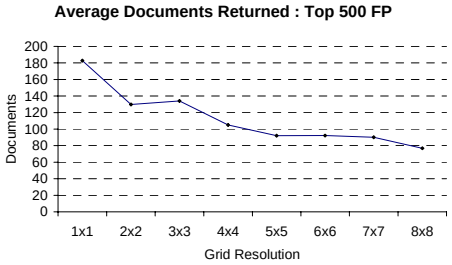

Figure 10. Average Documents Returned : Top 500 FP

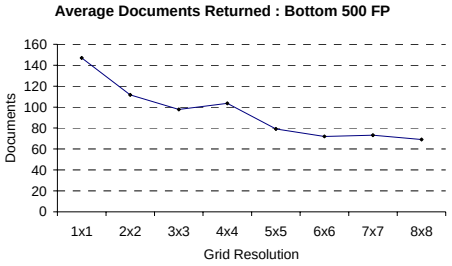

Figure 11. Average Documents Returned : Bottom 500 FP

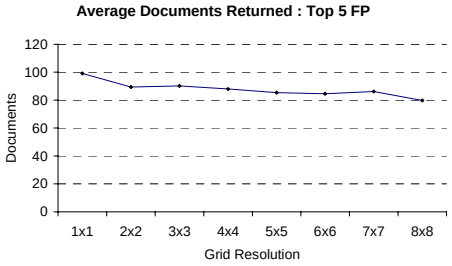

Figure 12. Average Documents Returned : Top 5 FP

6 Conclusions

Spatial indexing of web documents in combination with text indexing of the document content provides a means of managing and retrieving relevant web documents for purposes of geographic search that is superior to conventional text indexing alone. Effective spatio-textual indexing will help to ensure that all relevant documents are retrieved, even when they do not include geographical terms that match exactly with those in a user's query. Spatial indexing also facilitates processing search engine queries that include spatial relationships to a named place, such as *near* and *north of*. In this work three indexing schemes combining spatial and textual indexing have been presented and compared with each other and with a pure text index (PT), using a web collection of real documents that have been classified geographically with regard to their geographical context. The ST indexing scheme is spatial primary and creates a set of spatially-specific text indexes. The TS scheme is text primary and associates spatially ordered lists of documents with the indexed terms. The T scheme uses a pure text index to find relevant documents containing the non-geographical query terms and a separate spatial index of documents to find documents whose footprint intersects the query footprint, before intersecting the result sets.

In a comparison of the resulting index sizes, both ST and TS proved expensive relative to PT. The T scheme resulted in very little additional storage cost. The high storage overheads of ST and TS can be explained largely by the fact that the document footprints usually consist of many individual footprints (on average 21) reflecting the multiple places referred to in the document. This could be alleviated by more sophisticated geoparsing and geocoding procedures which identified a few dominant individual places to which a document refers as described in [1]. Query times for TS and ST were usually faster than for PT for all spatial index grid resolutions considered (note that PT has to process all query terms, whereas the other methods convert the geographical terms to a query footprint for access to the spatial elements of the indexes and do not use them for access to the text index component). An exception to this performance occurred with the ST index for queries with very

large query footprints. The T scheme produced slower query times on the whole than the other spatio-textual schemes but, for the most realistic query sets, this was about double, while in comparison with PT it was only about 25% greater. There is little doubt that the slower times reflect the pragmatic but inefficient use of unix script functions such as *grep*. It showed the same degradation with increasing grid resolution for the query set using very large query footprints. From the point of view of speed, the TS scheme was consistently advantageous. All spatio-textual schemes behaved the same in returning fewer documents with increasing spatial grid resolution. This reflected the closer approximation of the grid cells to the query footprint with increasing resolution.

It should be stressed that the objective of this study was to investigate the viability of spatio-textual indexing schemes in comparison with pure text indexing. It is assumed that spatio-textual indexing will retrieve more relevant documents (i.e. improve recall) in comparison with pure text methods, as it will be able to find documents referring to contained and nearby places to the geographical query place, and to places with alternative names to that specified in the query. In summary, the study has demonstrated that one scheme T introduced minimal storage overheads while resulting in only a small degradation in query times relative to PT, except for the case of very large query footprints. The TS scheme gave the most consistently good query time performance but was marred by the large storage overheads which could be improved by reducing the number of individual footprints per document.

There is clearly scope for further work to refine the methods described with regard to improved geo-tagging and improved document merging methods. It would also be appropriate to investigate higher spatial grid resolutions and other spatial indexing methods, as well as the use of a much larger web collection. It should be remarked that spatial index access times were not a significant overhead in these experiments and improved spatial indexing by itself could not be expected to result in great overall improvements. It is however of interest to investigate closer integration of text and spatial indexing, such as the use of spatial cell identifiers (locational keys) as part of the text index. A preliminary study which concatenated text with spatial keys, using simulated data, was described in [7].

An issue requiring further attention is that of user evaluation of the results. It has been stated that spatio-textual indexing is assumed to generate superior results relative to pure text indexing. Provided that the geoparsing and geocoding of documents is done effectively, i.e. documents are on the whole correctly categorized with regard to their geographical context, then this appears to be a reasonable assumption. Future studies will conduct such an evaluation to test this assumption when an adequate test collection becomes available.

Acknowledgements

This research was funded by the EC SPIRIT project IST-2001-35047 : Spatially-aware information retrieval on the internet.

References

1. Amitay, E., et al., 2004. Web-a-where: geotagging web content. in *27th ACM SIGIR Conference*: 273-280

2. Baeza-Yates, R. and B. Ribeiro-Neto, 1999. *Modern Information Retrieval*: Addison Wesley.
3. Buyukokkten, O., et al., 1999. Exploiting geographical location information of web pages. in *WebDB'99* (with ACM SIGMOD'99).
4. Cunningham, H., et al., 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. in *40th Anniversary Meeting of Assoc. for Computational Linguistics (ACL'02)*.
5. Ding, J., L. Gravano, and N. Shivakumar, 2000. Computing Geographical Scopes of Web Resources. in *26th Int. Conf. on Very Large Data Bases (VLDB)*, 545–556.
6. GoogleLocal. http://www.local.google.com
7. Jones, C.B, A.I. Abdelmoty, D. Finch, G. Fu and S. Vaid, 2004. The SPIRIT Spatial Search Engine :Architecture, Ontologies and Spatial Indexing. *in Third Int. Conf on Geographic Information Science GIScience 2004. LNCS* 3234, 125-39.
8. Jones, C.B., A.I. Abdelmoty, and G. Fu., 2003. Maintaining ontologies for geographical information retrieval on the web. in *On The Move to Meaningful Internet Systems 2003, ODBASE'03.LNCS* 2888, 934-51.
9. Jones, C.B., et al., 2002. Spatial information retrieval and geographical ontologies an overview of the SPIRIT project. in Proc ACM *SIGIR 2002*, 387-8.
10. Kornai, A. and B. Sundheim, eds. 2003. HLT-NAACL *Workshop on Analysis of Geographic References.*
11. Kreveld, M.van, I. Reinbacher, A. Arampatzis and R. van Zwol, 2004. Distributed Ranking Methods for Geographic Information Retrieval, in *Developments in Spatial Data Handling*, P.F. Fisher (Ed), Springer, 231-243.
12. McCurley, K.S., 2001. Geospatial mapping and navigation on the web. in *WWW10 Conference*, :http://www10.org/cdrom/papers/278/
13. Mirago. http://www.mirago.com
14. NorthernLight. http://www.northernlight.com
15. Purves, R. and C.B. Jones, 2004. *Workshop on Geographic Information Retrieval, SIGIR* 2004, http://www.sigir.org/forum/2004D/purves_sigirforum_2004d.pdf
16. Robertson, S.E. and S. Walker.1994 Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. *ACM SIGIR 1994*, 232-41.
17. Sagara, T. and M. Kitsuregawa, 2004 Yellow Page driven Methods of Collecting and Scoring Spatial Web Documents. in *SIGIR Workshop on Geographical Information Retrieval*, http://www.geo.unizh.ch/~rsp/gir/
18. Sanderson, M. and J. Kohler, 2004 Analyzing geographic queries. in *SIGIR Workshop on Geographic Information Retrieval*, http://www.geo.unizh.ch/~rsp/gir/
19. Silva, M.J., et al.2004 Adding Geographic Scopes to Web Resources. in *SIGIR Workshop on Geographical Information Retrieval*, http://www.geo.unizh.ch/~rsp/gir/
20. SPIRIT. http://www.geo-spirit.org/