# Ontology-Based Spatial Query Expansion in Information Retrieval

Gaihua Fu, Christopher B. Jones, and Alia I. Abdelmoty

School of Computer Science, Cardiff University, Cardiff, UK
{Gaihua.Fu, C.B.Jones, A.I.Abdelmoty}@cs.cf.ac.uk

**Abstract.** Ontologies play a key role in Semantic Web research. A common use of ontologies in Semantic Web is to enrich the current Web resources with some well-defined meaning to enhance the search capabilities of existing web searching systems. This paper reports on how ontologies developed in the EU Semantic Web project SPIRIT are used to support retrieval of documents that are considered to be spatially relevant to users' queries. The query expansion techniques presented in this paper are based on both a domain and a geographical ontology. The proposed techniques are distinguished from conventional ones in that a query is expanded by derivation of its geographical query footprint. The techniques are specially designed to resolve a query (such as *castles near Edinburgh*) that involves spatial terms (e.g. *Edinburgh*) and fuzzy spatial relationships (e.g. *near*) that qualify the spatial terms. Various factors are taken into account to support intelligent expansion of a spatial query, including, spatial terms as encoded in the geographical ontology, non-spatial terms as encoded in the domain ontology, as well as the semantics of the spatial relationships and their context of use. Some experiments have been carried out to evaluate the performance of the proposed techniques using sample realistic ontologies.

**Keywords:** Ontology, Semantic Web, Spatial Search, Query Expansion.

## 1 Introduction

The WWW holds vast amounts of information. However, users do not always get information they expect when searching the Web. One main reason for this is that existing web documents are rarely augmented with semantic annotation that describe their content, which would make them more easily accessible to automated search facilities. The Semantic Web is one of several proposed solutions to resolve this problem [29]. One aim of the Semantic Web is to enrich the current web documents with some well-defined meaning (meta-data), so that the existing web searching systems can be extended to have more advanced capabilities to find these resources more effectively. It has long been recognized in the Semantic Web research that ontologies play a key role as they can be used as a source of shared and precisely defined terms for such meta-data [14, 19].

Apart from annotating web documents with semantic information, ontologies have also been employed to resolve the mismatch problems between queries and

R. Meersman and Z. Tari (Eds.): CoopIS/DOA/ODBASE 2005, LNCS 3761, pp. 1466–1482, 2005.

documents, i.e. a query may not be expressed in terms that match the ones contained in some of the relevant documents. Traditionally this is dealt with using query expansion techniques which expand a query with the terms (as encoded in ontologies or other knowledge resources) that are considered to be related to the ones in the query, so that the relevant documents can be retrieved. Most of these studies use a term-based method [25, 1, 10, 30, 7]. For example, a query expansion method is introduced in [28] which extends a query with the words that are lexically related to the original query words using WordNet. A method is introduced in [15] to expand a query term with the ones that can be reached transitively in a concept network that is built up according to a thesaurus.

While these studies are useful for processing a general query, they provide inadequate support for processing a spatial query. A spatial query is different from a generic one in that it usually includes one or more spatial terms. It is often used by a user when he/she wishes to find Web resources that are related to a place. An example of such a query is *castles near Edinburgh*. Support for this type of query is necessary as most human activities are rooted in geographical space in some aspect, and therefore many documents include references to geographical context, typically by means of place names. Conventional search engines treat spatial terms involved in a query in the same way as other terms and can not always ensure good search results due to the lack of spatial awareness. This has led to research interests in developing spatial search techniques to help users find resources in which the subject matter is related to a place [13, 12, 22].

As with a generic query, there is also a need to expand a spatial query. While query expansion has been studied extensively in the literature, the interest here is how to expand a spatial query so that documents that are considered to be *spatially* relevant can be retrieved. A document can be spatially relevant to a query in different ways. It may be spatially relevant to a query by involving a geographical term that is considered to be an alternative name for the one appearing in the query. A document may also be spatially relevant to a query by involving places which satisfy the specified spatial relationship with the one appearing in the query. An example of this is a query looking for *castles near Edinburgh*. The relevant documents may not only include the ones that describe castles in *Edinburgh*, but also the ones that describe castles in places such as *West Lothian* and *Midlothian*, which are geographically *near* to *Edinburgh*.

Conventional term-based query expansion techniques can be utilised to resolve a spatial query. However, the danger is that they may introduce too many query terms in spatial context, perhaps many thousands, and may therefore become intractable for the query processing facilities. Another challenge in dealing with spatial query expansion is that a spatial relationship involved in a query can be vague. Its interpretation can vary with respect to different users' intentions, as well as depending on the types of spatial and non-spatial terms involved in a query. For example, one user may use *near* to refer to places that are either inside of or adjacent to a place presented in a query, and another user may use it to refer to places that are only adjacent to the specified place. Also, a spatial

relationship may need to be interpreted differently due to different subject matters involved. For example, *near* in the query *lakes near Edinburgh* may need to be treated differently from *near* in *hotels near Edinburgh*.

In this paper we report the spatial query expansion techniques developed in the EU Semantic Web project SPIRIT. The query expansion techniques presented in this paper are based on both a domain and a geographical ontology. Different from term-based query expansion techniques, the proposed techniques expand a query by trying to derive its geographical query footprint, and it is specially designed to resolve a spatial query. Various factors, such as types of spatial terms as encoded in the geographical ontology, types of non-spatial terms as encoded in the domain ontology, the semantics of the spatial relationships, their context of use, and satisfiability of initial search result, are taken into account to support expansion of a spatial query. The proposed techniques support the intelligent, flexible treatment of a spatial query when a fuzzy spatial relationship is involved. Some experiments have been carried out to evaluate the performance of the proposed techniques using sample realistic ontologies.

The remaining part of the paper is organized as following. Section 2 studies related work. Section 3 introduces the background knowledge of this research, discusses various factors that affect spatial query expansion, and presents how SPIRIT ontologies are designed to support spatial query expansion. Section 4 presents our method that supports spatial query expansion. Section 5 reports our experimental results. Section 6 concludes the paper and points out the possible future research.

## 2 Related Work

Query expansion is traditionally considered as a process of supplementing a query with additional terms as the assumption is that the initial query as provided by the user may be an inadequate representation of the user's information needs [28, 30, 15, 5, 7]. Query expansion techniques can broadly be classified in two categories: those based on the search results and those that are based on some forms of knowledge structure. The former group of techniques depends on the search process and uses relevance feedback in an earlier iteration of search as the resource to identify the query expansion terms [1, 4, 7]. The latter group of techniques is independent of the search process and additional query terms are derived by traversing a semantic network built up according to a knowledge structure. Knowledge structures used by this group of techniques can either be a general-purpose ontology (or thesaurus) [28], or an ontology built for a specific domain [15], or an ontology constructed from document collection based on the term clustering [20]. Work that combines the two approaches is reported in [30], where authors apply term clustering techniques to the local set of documents.

The work reported in this paper belongs to the second group of research, i.e., both a domain ontology and a geographical ontology are utilised to support query expansion. In the literature, there are several search engines that employ ontologies to support spatial query expansion [22, 18]. For example, Mirago has

developed a regional web search facility that provides spatial search services for several European countries including UK, Germany, France and Spain [22]. A user can issue a spatial query by typing a domain term and selecting from available place names (as encoded in a geographical ontology) the one that he/she would like his/her search to focus on, and documents that employ both the domain term and the spatial term in their text are retrieved. Mirago supports some limited spatial expansion by using the spatial containment relationship existing between places (as encoded in the geographical ontology). That is, if no or few documents are found according to a spatial query term, the term is replaced with a place name whose region immediately contains it.

In addition to term-based spatial query expansion research, recently some geographical search systems employ footprint-based spatial query expansion techniques to assist with retrieval of spatially relevant documents (the footprint of a query refers to the spatial search space of a query). For example, the geographical search engine developed by Vicinity [27] allows 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. Google has recently introduced a locational web search system based in the USA [13]. 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. Other research which considers the spatial search is that of [8, 2, 9, 3, 21]. All these spatial search engines support the *inside* spatial relationship, and a few of them support the *distance* relationship as well. Though relatively little has been published on the technology that underlies spatial query expansion by these systems, according to authors' investigation of some search results of these systems, it appears they perform query expansion by simply translating a place name into its corresponding coordinate footprint.

The main advantage of footprint-based query expansion is that it avoids introducing too many query terms, which, as discussed in [28], is not as effective as supposed to be. Furthermore, footprint-based query expansion can effectively avoids retrieval of irrelevant documents due to name sharing (according to [24], about 16.6 percent of European place names have multiple uses), which is usually inevitable in term-based query expansion. Finally, footprint-based expansion allows us to perform more accurate spatial relevance calculation by analysing the query footprint and the document footprint, which is not possible with term-based expansion.

The work reported in this paper studies footprint-based spatial query expansion techniques. It is distinguished from previous research in several aspects. First, it supports spatial query expansion especially when a fuzzy spatial relationship term such as *near* is presented in a query, which is largely not considered in other research. A wide range of spatial fuzzy spatial relationship terms are supported by the techniques proposed in this paper. Secondly, the proposed techniques support intelligent and flexible spatial query expansion. This is achieved by taking into account of various factors, e.g. spatial query term, non-spatial query terms, the use context of a spatial relationship etc., when computing a

query footprint. Thirdly, we support iterative spatial query expansion, i.e. a query footprint will be progressively extended when initial search results are not sufficient. This in one aspect ensures search satisfactory. On the other hand, it ensures the most spatially relevant documents will be retrieved first, which is difficult to be achieved with traditional query expansion techniques if not impossible.

## 3 SPIRIT Queries, Query Expansion and Ontologies

The work reported in this paper is part of the SPIRIT project (Spatially-Aware Information Retrieval on the Internet) [6]. The aim of SPIRIT is to develop Web search technology that is specialised for access to documents relating to places or regions referred to in a query. A primitive spatial query in SPIRIT can be formalised as a triple:

$$\langle what, rel, where \rangle$$

where the *what* term is used to specify a general non-spatial object, which may correspond to a physical or an abstract subject or activity; *where* is used to specify a spatially referenced term; the *rel* term is a spatial relationship which relates *what* and *where*.

The following concepts are used throughout the paper to illustrate our techniques. A spatial term is the one which has a footprint *P-footprint*.

**Definition 1.** *The footprint P-footprint of a spatial term indicates the geographical location of the intended place, and is specified in terms of map coordinates with a selected reference system.*

A document may have footprint *D-footprint* if it involves one/more spatial terms.

**Definition 2.** *A document footprint D-footprint defines the geographical coverage of a specified document, and it may consist of multiple P-footprints if more than one place name appears in the document.*

Given a spatial query $\langle what, rel, where \rangle$, the purpose of spatial query expansion in this research is to generate a query footprint (denoted as *Q-footprint*). Ideally, *Q-footprint* should be computed in such a way so that spatially relevant documents of $\langle what, rel, where \rangle$ are those whose document footprints fall in *Q-footprint*.

**Definition 3.** *A query footprint Q-footprint defines a geographical space that covers the intended spatial search extent of* $\langle what, rel, where \rangle$*, and it is specified in the form of map coordinates.*

Given $\langle what, rel, where \rangle$, deriving *Q-footprint* will start with the *P-footprint* of *where*. The most important information that influences *Q-footprint* is the *rel* term, and it determines what geographical area should be covered by *Q-footprint*.

For example, if *rel* is *near*, the query footprint may be assumed to be the area surrounding *where*. If *rel* is *north*, the geographical area that covers north of *where* should be returned.

Most spatial relationships are fuzzy, and their semantics can vary when used with different combinations of *what* and *where*. Consequently, *Q-footprint* may be different when the same *rel* is used in different contexts. Given $\langle what, rel, where \rangle$, we consider that the interpretation of *rel* is mainly determined by the following factors:

- the *type* of *where*. This is because the search extent is usually assumed differently where different types of *where* are presented in queries. For example, given *near*, we tend to assume a bigger search space when *where* is of type *city* than when it is of type *village*.
- the *P-footprint* of *where*. Some places are of the same type, but the areas they cover can vary. For example, both *London* and $Cardiff$ are type of *city*, but the area of *London* is much bigger than that of $Cardiff$. Therefore it is reasonable to assume a larger neighbourhood region of *London* than that of *Cardiff*.
- the *what* term. Given a geographical area, the distribution densities of different *what* subjects may vary, and therefore some subjects may have more documents describing them than others. For example, there are more *hotels* than *airports* for most places. For a subject which has a sparse distribution density in an area, it tends to require a bigger search space in order to find some relevant documents.
- the user's intention of using a *rel* term. Different users may employ a same *rel* with different meanings in mind. For example, one user may use *near* to refer to a region that covers both the *where* and its surrounding areas, while another user would use *near* only to refer to the neighbouring regions of *where*. Therefore it is desirable that *rel* can be interpreted by taking into account of the user's intention in mind.

We are aware that other factors may also affect the interpretation of a *rel* term, for example, the *population* of *where* if it is an inhabited area. However, most of these factors apply to the specific type of queries (e.g. *population* only needs to be considered for a query whose *where* term represents an inhabited place), whereas the factors considered in this research apply to generic spatial queries. Therefore our techniques support spatial query expansion by mainly considering the generic factors. However, when a query footprint does not produce good search results, our techniques support iterative spatial query expansion (see Section 4 for details).

To support spatial query expansion, the SPIRIT system has incorporated into its architecture an ontology component, of which the primary parts are a domain ontology and a geographical ontology (or geo-ontology)[1]. The domain

[1] SPIRIT ontology design was also driven by other spatial search requirements, e.g. spatial query disambiguation, spatial relevance ranking, spatial index and annotation of web resources, as discussed in [11, 16].

ontology models the terminologies of one application area or domain, and is used to resolve the *what* aspect of a SPIRIT query. Modelling of domain-specific terminology is accomplished using conventional thesauri methods. Equivalent terms or synonyms are represented via `USE` and `USE-FOR` relations. Hierarchical relations whether generic (is-a) or metronymic (part-of) are represented with Broader Term (`BT`) and Narrower Term (`NT`) relations. For each term, the domain ontology maintains a coefficient that indicates the influence of it on the interpretation of a spatial relationship, and this is derived by carrying out some document density studies.

The *where* aspect of the SPIRIT query is dealt with the SPIRIT geo-ontology, which is constructed to provide a knowledge structure of the interested geographic space. Several types of information are encoded in the geo-ontology, including the various names that a place is known by, the place types with which it can be categorised, its topological relationships (such as *partof* and *containing*) with other places, and its geographical footprints (*P-footprint*). For each category of place, the geographical ontology maintains a coefficient that indicates the influence of it on the interpretation of a spatial relationship, and this is derived by carriying out some user studies.

## 4 A Method for Deriving Spatial Query Footprint

This section describes how spatial query expansion is performed by employing the SPIRIT ontologies. The proposed techniques are mainly designed to handle spatial queries with fuzzy spatial relationships presented, and the group of spatial relationships that can be handled by using techniques proposed in this paper includes *in, near, outside, north-of, south-of, east-of, west-of* and *within a specified distance*[2].

Apart from a domain ontology and a geographical ontology, we assume the availability of the alternative interpretation of a spatial relationship *rel*. For example, for *near*, three options may be available for its interpretation: an area covers only *where*, an area covers both *where* and its surrounding regions and an area covering neighbouring regions of *where*. The statistical data of a spatial relationship in search needs to be maintained to record the option that a user may choose in search processes, and the frequency that an option is chosen for interpreting a spatial relationship [3].

The proposed techniques support iterative spatial query expansion. This is necessary for several reasons. First, for some topics, inadequate documents may exist on Web to describe them. Secondly, some information encoded in the ontologies, such as coefficient data which indicates the influence of a domain term on the interpretation of a *rel* term, may not be as valid as they are supposed to be, especially when experiments for obtaining these parameter values are too expensive to perform exhaustively. Finally, query footprint will be derived by

[2] Other spatial relationships need to be treated differently and our follow up paper will elaborate on this.

[3] This is achieved by maintaining a log file.

taking some generic factors into accounts, while some specific types of query may need to consider other factors, as stated in Section 3. Therefore it is desirable that that spatial query expansion can be performed iteratively when initial search results are not satisfactory.

In what follows, we will use $Q\text{-}footprint_i$ to denote the query footprint generated at the *i-th* iteration of query expansion. We first describe how the initial query footprint $Q\text{-}footprint_1$ is computed, and then describe how query footprint can be incrementally expanded when initial search results are inadequate. We will use the geographical space (which covers the UK county "Oxfordshire" and its surrounding area) shown in Figure 1 to illustrate the techniques proposed.

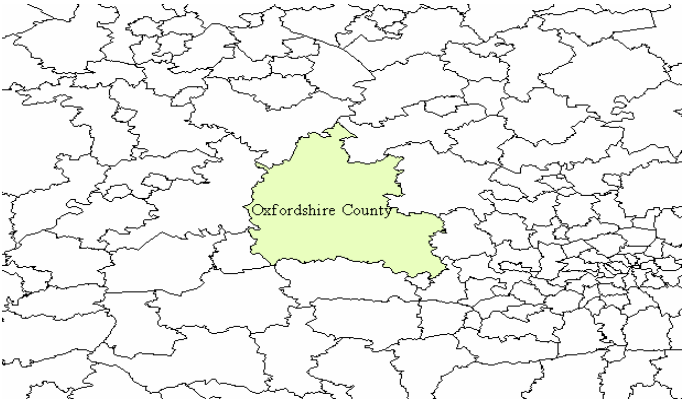


**Fig. 1.** `Oxfordshire` and its Surrounding Area

### 4.1 Initial Spatial Query Expansion

The following steps describe how $Q\text{-}footprint_1$ is generated.

1. Though *P-footprint* of the *where* term is the starting point from which $Q\text{-}footprint_1$ is generated, the type of geometric operation that is performed over *P-footprint* for generating $Q\text{-}footprint_1$ is determined by *rel*. For example, if *rel* is *near*, then a buffer operation needs to be performed over *P-footprint* for generating $Q\text{-}footprint_1$. Therefore the first step of computing *Q-footprint* is to determine the type of geometric function required according to *rel*. This is shown by using following function:

$$GeoOp = \beta(rel) \tag{1}$$

where the function $\beta$ maps a spatial relationship *rel* to a corresponding geometric function name. For example, if *rel* is *near*, the function $\beta$ will generate value $Buffer$ for *GeoOp*. Different *rel* terms result in query footprints of different orientation and geometries. Figure 2 shows some example query footprints (polygons plotted with bold lines), when *rel* stands for *near, outside-of* and *north-of*. When a query is in the form of $\langle what, near, Oxfordshire \rangle$,

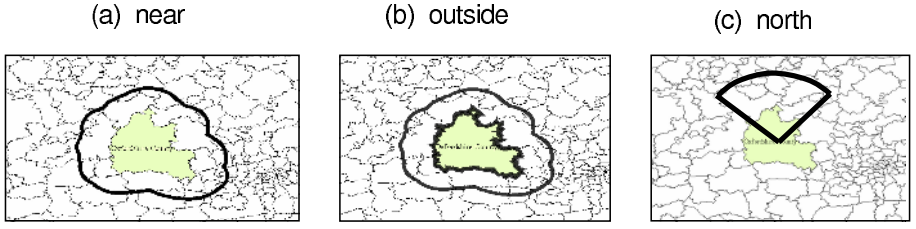


**Fig. 2.** Different *rel* Terms Resulting in Different Query Footprints

*Q-footprint* is the space that covers both $Oxfordshire$ and its surrounding areas. *Q-footprint* for $\langle what, outside\text{-}of, Oxfordshire\rangle$ is quite similar to the one for $\langle what, near, Oxfordshire\rangle$, but it only covers surrounding regions of $Oxfordshire$. If a query is in the form of $\langle what, north\text{-}of, Oxfordshire\rangle$, then the area that covers the north and northern part of $Oxfordshire$ is returned as *Q-footprint*.

2. To derive exact geographical coverage of $Q\text{-}footprint_1$, a geometric function *GeoOp* requires the following parameters:
   (a) the *P-footprint* of the *where* term, and it can be retrieved from the SPIRIT geo-ontology. This gives us the initial geometry from which $Q\text{-}footprint_1$ is to be generated;
   (b) a geometric distance $d$ that is required for extending *P-footprint* to generate $Q\text{-}footprint_1$. The group of fuzzy *rel* terms studied in this paper determines that $Q\text{-}footprint_1$ is generated by extending *P-footprint* at a specified distance in a certain way. For example, if *rel* is *near*, $Q\text{-}footprint_1$ may be generated to cover areas extended from *P-footprint* at a specified distance. If *rel* is *north*, $Q\text{-}footprint_1$ may be generated to cove areas extended from the north part of *P-footprint* at a specified distance. The exact distance $d$ for geometric expansion is determined by the following:
      i. the area size of *P-footprint*, and it is used to determine the initial extension distance using the formula shown below:[4]

$$id = \sqrt{\frac{area(P\text{-}footprint)}{\pi}} \tag{2}$$

      That is, the initial extension distance is assigned the approximate radius of *P-footprint*;
      ii. a coefficient $p_1$ which determines the influence of the *what* term, this can be retrieved from the SPIRIT domain ontology;
      iii. a coefficient $p_2$ which determines the influence of the *where* term, that can be retrieved from the SPIRIT geo-ontology;

   The exact expansion distance is therefore determined by the following:

$$d = id * p_1 * p_2 \tag{3}$$

[4] This formula is used in our preliminary study, and some more user and performance experiments need to be carried out to validate it.

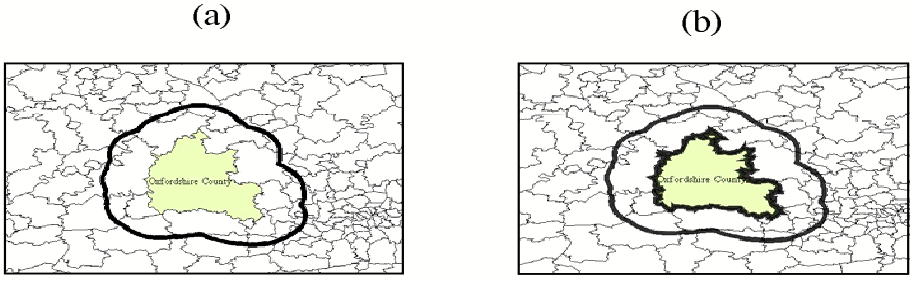


**Fig. 3.** Different Interpretation of Spatial Relationship *near*

(c) As we mentioned earlier, each *rel* may have different interpretations. This can either be chosen by a user or be derived from the SPIRIT log file which encodes the most frequently used option for a specified *rel*. This is assigned to parameter $p_3$[5]. For example, Figure 3 shows query footprints when *near* is interpreted differently – one covers both the *where* and its surrounding areas and another just covers the neighbouring regions of *where*.

3. The parameters derived in step 2(a), 2(b) and 2(c) are passed on to the geometric function *GeoOp* generated in Step 1 to derive *Q-footprint*$_1$.

$$\textit{Q-footprint}_1 = \textit{GeoOp}(\textit{P-footprint}, d, p_3) \tag{4}$$

### 4.2 Iterative Spatial Query Expansion

When an initial search results fail to satisfy the user's query need, our method regenerates *Q-footprint* to cover some regions beyond that of *Q-footprint*$_1$. This section shows how this is achieved. Given $\langle what, rel, where\rangle$, we derive *Q-footprint*$_i$ if the search results of *Q-footprint*$_1, \ldots,$ *Q-footprint*$_{i-1}$ are not satisfactory[6]. The procedure below describes how iterative spatial query expansion is performed:

1. derive *Q-footprint*$_i$ if the iteration criterion is satisfied (e.g. when search result is not satisfactory after *i-1* rounds of iteration). *Q-footprint*$_i$ can be derived largely using spatial query expansion procedure described in Section 4.1. The difference is that we further enlarge the geometric distance $d$ generated in the formula (3) according to:

$$d = d * i \tag{5}$$

[5] That is, the most frequently used interpretation of *rel* is used by the system by default. However, the SPIRIT user interface allows a user to choose other options as well.

[6] Various factors can control iterative query expansion process, e.g. a new iteration can be triggered when no or few documents are retrieved in initial search, and iteration can be interrupted if the allocated search time runs out. This is beyond the topic of this paper and therefore will not be discussed further here.

2. subtract from $Q\text{-}footprint_i$ the area covered by $Q\text{-}footprint_1, \ldots, Q\text{-}footprint_{i-1}$ so that to avoid spatial search redundancy:

$$Q\text{-}footprint_i = Q\text{-}footprint_i - \sum_{1}^{i-1} Q\text{-}footprint_k \tag{6}$$

Figure 4 shows query footprints that are progressively generated in order to find documents for the query $\langle airports, near, Oxfordshire \rangle$, and we can see that it has been spatially expanded three times in its effort to find spatially relevant documents.

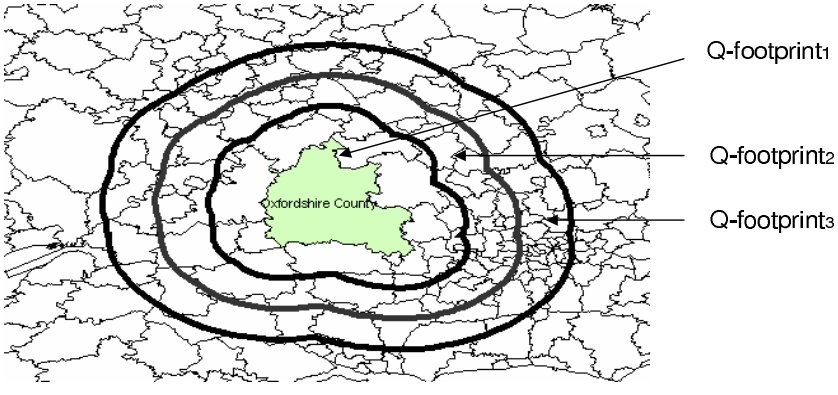


**Fig. 4.** Iterative Spatial Query Expansion

## 5 Implementation and Evaluation

To verify the spatial query techniques proposed, we have carried out some experiments. In this section, we will demonstrate how query expansion techniques are used in SPIRIT to improve search results, we also report on the experiments which were carried out to study the time cost for performing spatial query expansion using SPIRIT ontologies.

Query expansion techniques are implemented using Java, and they interact with SPIRIT ontology databases (composed of a domain ontology and a geo-ontology) to compute query footprints. The domain ontology contains the terms that are used in tourism area, and 2223 terms are encoded. The geo-ontology contains geographical places of several European countries, including United Kingdom, France, Germany and Switzerland, and 125,812 places are encoded. Both domain and geo-ontology are stored in Oracle 9.2.0. Once a query footprint is generated, it is feed to SPIRIT search component to retrieve the relevant documents. All experiments were carried out on a Pentium 4 PC with a 2.00 GHz processor and 516 MB of memory, running Microsoft Windows/XP. The SPIRIT adopts a distributed architecture (see [17] for details), and query expansion services talks to other components of the system through Apache SOAP.

### 5.1 Precision Study

This section demonstrates the effectiveness of the spatial query expansion techniques proposed. This is achieved by comparing the search results obtained by the SPIRIT system when spatial query expansion option is either switched on and off. When spatial query expansion is on, the SPIRIT system performs query expansion using techniques proposed, search is carried out basing on both the spatial and the textual index of web collection, and relevant ranking is performed using techniques proposed in [26]. When spatial query expansion is off, all query terms (including spatial and non-spatial ones) are send to the search component to perform a textual based search, and `BM25` proposed in [23] is used to rank the search results.

**Table 1.** Search Topics

| | |
|---|---|
| query 1 | ⟨castles, inside, Cardiff⟩ |
| query 2 | ⟨castles, near, Cardiff⟩ |
| query 3 | ⟨castles, north-of, Cardiff⟩ |
| query 4 | ⟨castles, outside-of, Cardiff⟩ |

The experiments were carried out using a set of queries ( shown in Table 1). The queries involve *rel* terms *inside, near, north-of* and *outside-of*. Since other *rel* terms such as *south-of* and *east-of* are treated similarly with *north-of* using our techniques, we consider the set of *rel* terms are sufficient for evaluation purposes. The results produced from running these queries were analysed for P10 (precision at 10) accuracy. The top ten results were examined by human users to judge their spatial relevance to the given queries. To help with judging spatial relevance of the retrieved documents, the UK city *Cardiff* and its surrounding areas, which are familiar to the intended users, were chosen for the queries to focus on.

A retrieved document was classified as three types in our experiments: relevant, irrelevant, and partially relevant. The first two types are easy to understand. A document is classified as *partially relevant* it is not designed to describe the search topic but it contains a link that points to a relevant page, e.g. a directory page. We note that a *rel* term can be interpreted differently using our system, however due to human effort required, we were only able to perform experiments for a fixed number of settings. Table 2 shows the experiment results, where columns 3, 4, 5 display the numbers of relevant, partially relevant and irrelevant documents retrieved.

When *rel* term is *inside*, the query footprint *Q-footprint* is *P-footprint* of *where*. It is not obvious that the search system performed better when spatial query expansion option was switched on. However, with query expansion option switched on, we observed that documents, which describe castles in terms of subareas of *Cardiff*, or alternative names of *Cardiff*, i.e. *Caerdydd*, were retrieved. This did not happen when query expansion option was switched off. The main reason for this is that our spatial query expansion is footprint-based, and retrieved documents are the ones whose documents footprints fall in query

**Table 2.** Experimental Results

| query | spatial query expansion | relevant | partially relevant | irrelevant |
|---|---|---|---|---|
| 1 | off | 3 | 6 | 1 |
| | on | 2 | 6 | 2 |
| 2 | off | 1 | 3 | 6 |
| | on | 5 | 5 | 0 |
| 3 | off | 2 | 3 | 5 |
| | on | 4 | 6 | 0 |
| 4 | off | 0 | 4 | 6 |
| | on | 5 | 4 | 1 |

footprint. Different documents may have different geographical terms in their text, but if these geographical terms refer to same places, these documents have the same document footprints, which all fall in *Q-footprint*. Documents specified in term of subareas of *where* have footprints which are subsets of *Q-footprint*, therefore are retrieved as well.

When *rel* term is *near*, *Q-footprint* were generated covering *P-footprint* of *where* plus its surrounding areas. The search system performed better with spatial query expansion switched on – the top 10 retrieved documents are either relevant or partially relevant. The footprint-based query expansion enabled us to retrieve documents which describes castles not only in $Cardiff$ but also in places like *Caerphilly, Newport, Dinas Powys, Abergavenny* and *Swansea*. Since these places are geographically close to *Cardiff*, the retrieved documents are spatially relevant to the query. When spatial query expansion was off, it appeared that all retrieved documents involve the terms *castles, near* and *Cardiff*. Unfortunately, many of these documents do not actually describe castles *in* or *near* to $Cardiff$.

When *rel* term is *north-of*, *Q-footprint* were generated covering northern part of *where* plus areas that are north of *where*. From Table 2, we can see that the system performed considerably better when spatial query expansion was switched on. The reason is the same with *near* – the footprint-based query expansion enables us to retrieved documents whose footprints satisfy specified geometric relationship with query footprint. However, when spatial query expansion was off, many documents retrieved are the ones which happen to have the terms *castles, north-of* and *Cardiff* presented, but do not actually describe castles in the northern or north of *Cardiff*.

When *rel* term is *ouside-of*, *Q-footprint* were generated covering only surrounding area of *where*. Same with *near* and *north-of*, the system presented its inability to deal this type queries when spatial query expansion was off, whereas it performed considerably better when spatial query expansion option was on.

### 5.2 Time Cost Study

Due to the complexity of the original geometric footprint of a place, only two approximation representations of a footprint, MBR and convex hull polygon,

were utilised to deal with spatial query expansion in SPIRIT. An MBR is the minimum bounding rectangle of a geometry object, and a convex hull polygon is the smallest convex polygon that completely encloses a geometry object.

We first compared the time costs of query expansion by using MBRs and convex polygons, and the mean response time of using two types of footprint for query expansion is shown in Figure 5.

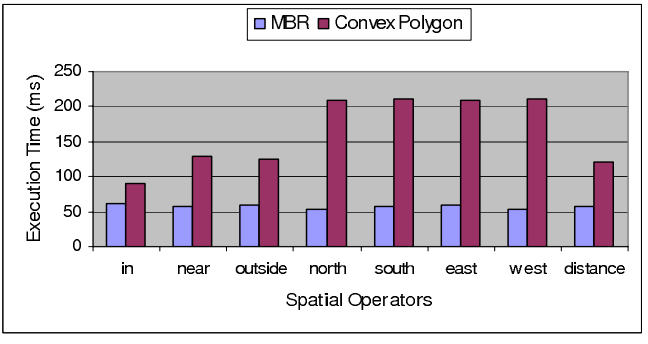


**Fig. 5.** Response Time of Query Expansion by using MBRs and Convex Polygons

From Figure 5, we can observe that it requires more CPU time to derive query footprint using convex polygon than using MBR. This is mostly due to the complex nature of convex polygons. A MBR is composed of two coordinate points, while a convex polygon can have more coordinate points, ranging from 7 to 38 according to our geographical ontology. However, the CPU time required for deriving query footprints using both MBR and convex polygon are in a range that is acceptable in a SOAP-based distributed search environment, i.e. about 60 milliseconds for MBR and about 210 milliseconds for convex polygon.

We then studied the time cost of query expansion using convex polygons with different complexity, i.e. convex polygons composed of different numbers of coor-

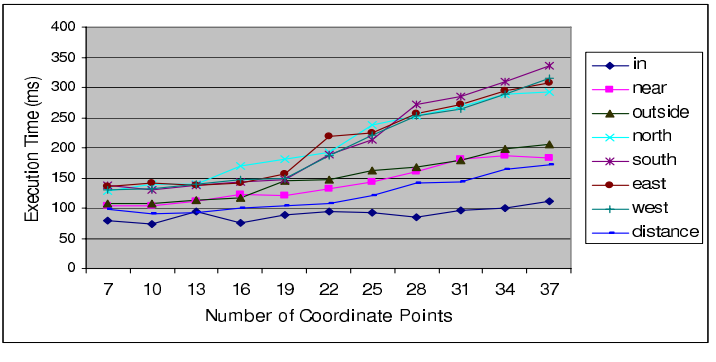


**Fig. 6.** Impact of Coordidnate Point Number

dinate points, and the result is shown in Figure 6. As we can see, that response time increases with number of coordinate points – the more coordinate points a convex polygon has, the more CPU time is required for deriving the query footprint. This increase is obvious when dealing with spatial relationships *north, south, east* and west. However, for all spatial relationship terms, the increase displays a linear tendency.

## 6 Conclusions

In this paper we have introduced an ontology-based spatial query expansion method that supports retrieval of documents that are considered to be spatially relevant. The proposed method expands a spatial query by trying to derive its geographical query footprint, and it is specially designed to resolve a query that involves a fuzzy spatial relationship. Both a domain and a geographical ontology are employed to support spatial query expansion. Various factors are taken into consideration for supporting intelligent expansion of a spatial query, and proposed method also supports iterative spatial query expansion when initial spatial searches are not satisfactory. Our experiments show that the proposed method can considerably improve search results when a query involves a fuzzy spatial relationship, and experiments also show that proposed method works efficiently using realistic ontologies in a distributed spatial search environment. The method reported in this paper is proposed to deal with a group of spatial relationships that frequently appear in spatial search, and how to resolve other spatial relationships, e.g. *between*, still requires further investigation.

## Acknowledgement

This work is funded by Grant IST-2001-35047 from EC Fifth Framework Programme.

## References

1. R. Attar and A. S. Fraenkel. Local Feedback in Full-Text Retrieval Systems. *Journal of the ACM*, 24(3):397–417, July 1977.
2. S. Bressan, B. 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.
3. 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*. ACM press, 1999.
4. D. Cai, C. J. Rijsbergen, and J. M. Jose. Automatic Query Expansion based on Divergence. In H. Paques, L. Liu, and D. Grossman, editors, *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM-01)*, pages 419–426, New York, Nov. 5–10 2001. ACM Press.

5. C. Carpineto, R. de Mori, G. Romano, and B. Bigi. An Information-Theoretic Approach to Automatic Query Expansion. *ACM Transactions on Information Systems*, 19(1):1–27, 2001.
6. C.B. Jones and R. Purves and A. Ruas and M. Sanderson and M. Sester and M.J. van Kreveld and R. Weibel. Spatial Information Retrieval and Geographical ontologies: an Overview of the SPIRIT Project. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 387–388, 2002.
7. H. Cui, J. Wen, and M. Li. A Statistical Query Expansion Model Based on Query Logs. *Journal of Software*, 14(9):1593–1599, 2003.
8. D. Egnor. http://www.google.com/programming-contest/winner.html.
9. 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.
10. E. N. Efthimiadis. Query Expansion. In M. E. Williams, editor, *Annual Review of Information Science and Technology*, volume 31, pages 121–187. American Society for Information Science, 1996.
11. G. Fu, C. Jones, and A. I. Abdelmoty. Building a Geographical Ontology for Intelligent Spatial Search on the Web. In *Proceedings of IASTED International Conference on Databases and Applications*, pages 167–172. Spriner Verlag, 2005.
12. GBdirect Ltd. SomeWhere Near. http://somewherenear.com/.
13. Google. Google Location Search. http://local.google.com/lochp.
14. I. Horrocks and P. F. Patel-Schneider. Reducing OWL entailment to description logic satisfiability. In D. Fensel, K. Sycara, and J. Mylopoulos, editors, *Proc. of the 2003 International Semantic Web Conference (ISWC 2003)*, number 2870 in Lecture Notes in Computer Science, pages 17–29. Springer, 2003.
15. K. Järvelin, J. Kekäläinen, and T. Niemi. ExpansionTool: Concept-Based Query Expansion and Construction. *Information Retrieval*, 4(3/4):231–255, 2001.
16. C. Jones, A. Abdelmoty, and G. Fu. Maintaining Ontologies for Geographical Information Retrieval on the Web. In *Proceedings of OTM Confederated International Conferences CoopIS, DOA, and OOBASE*, pages 934–951. Spriner Verlag, 2003.
17. C. Jones, A. I. Abdelmoty, D. Finch, G. Fu, and S. Vaid. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In *Proceedings of the 3rd International Conference on Geographic Information Science*, pages 125–139.
18. C. Jones, D. Tudhope, and H. Alani. Augmenting Thesaurus Relationships: Possibilities for Retrieval. *Journal of Digital Information*, 1(8), Jan. 15 2001.
19. O. Lassila and R. R. Swick. Resource description framework (rdf) model and syntax specification. W3C Recommendation, 1999. Available at http://www.w3.org/TR/1999/REC-rdf-syntax-19990222.
20. R. Mandala, T. Tokunaga, and H. Tanaka. Combining General Hand-Made and Automatically Constructed Thesauri for Query Expansion in Information Retrieval. In D. Thomas, editor, *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI-99-Vol2)*, pages 920–925, S.F., July 31–Aug. 6 1999. Morgan Kaufmann Publishers.
21. K. McCurley. Geospatial Mapping and Navigation of the Web. In *Proceedings of Tenth International World Wide Web Conference*, page Session P7. ACM press, 2001.
22. Mirago: Mirago the UK Search Engine. http://www.mirago.co.uk/.
23. S. E. Robertson, S. Walker, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference (TREC3)*.

24. D. A. Smith and G. S. Mann. Bootstrapping Toponym Classifiers. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 45–49.
25. K. Sparck Jones. *Automatic Keyword Classification and Information Retrieval*. Butterworths, London, 1971.
26. M. van Kreveld, I. Reinbacher, A. Arampatzis, and R. van Zwol. Distributed Ranking Methods for Geographic Information Retrieval. In *Proceedings of 11th Int. Sympos. on Spatial Data Handling: Developments in Spatial Data Handling*, pages 231–243.
27. Vicinity.com. http://home.vicinity.com/us/mappoint.htm.
28. E. M. Voorhees. Query Expansion Using Lexical-Semantic Relations. In W. B. Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 61–69. ACM/Springer, 1994.
29. W3C. Semantic Web. http://www.w3.org/2001/sw/, 2004.
30. J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11. ACM Press, 1996.