Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text

Ross S. Purves
University of Zurich
ross.purves@geo.uzh.ch

Paul Clough
University of Sheffield
p.d.clough@sheffield.ac.uk

Christopher B. Jones
Cardiff University
JonesCB2@cardiff.ac.uk

Mark H. Hall
Martin Luther University Halle-Wittenberg
mark.hall@informatik.uni-halle.de

Vanessa Murdock
Microsoft
vanmur@microsoft.com

February 9, 2018

Abstract

Significant amounts of information available today contain references to places on earth. Traditionally such information has been held as structured data and was the concern of Geographic Information Systems (GIS). However, increasing amounts of data in the form of unstructured text are available for indexing and retrieval that also contain spatial references. This monograph describes the field of Geographic Information Retrieval (GIR) that seeks to develop spatially-aware search systems and support user's geographical information needs. Important concepts with respect to storing, querying and analysing geographical information in computers are introduced, before user needs and interaction in the context of GIR are explored. The task of associating documents with coordinates, prior to their indexing and ranking forms the core of any GIR system, and different approaches and their implications are discussed. Evaluating the resulting systems and their components, and different paradigms for doing so continue to be an important area of research in GIR and are illustrated through a number of examples. The article concludes by setting out a range of future challenges for research in this field.

1 Introduction

1.1 Setting the scene

The importance of location in search seems obvious. A large proportion of search queries include explicitly geographic search terms, for example in the form of place names (Gan *et al.*, 2008; Aloteibi and Sanderson, 2014). Local search, that is the provision of access through search engine interfaces to structured information, such as opening hours, business

locations or local product availability, is estimated to be accessed by 80% of search engine users, and these users actively wish advertising to be locally relevant to their needs[1]. Location-based services, where a user's current or predicted location is used as real time contextual information in the delivery of services, are propagating at a furious pace, with a focus on providing information relevant to mobile users' needs (Reichenbacher *et al.*, 2016).

Although search engines have invested heavily in local search in recent years, results retrieved by some search engines are mostly limited to information found in commercial directory listings. The situation is certainly improving with the increasing availability via local search of other non-commercial structured or semi-structured georeferenced sources and their associated web sites. There is however a large body of unstructured web content that refers to geographical information but which at present will only be retrieved if there is a direct match between the query terms and terms in the document. Effective access to unstructured documents, in which geographical relevance can be inferred, requires methods that can recognise the presence of geographic references in documents and resolve these unambiguously to locations on the earth's surface. This includes automated interpretation both of place names and of qualifying spatial relationships in queries and in documents. The development of such methods in combination with techniques for indexing, ranking and retrieval of the associated content is the focus of this article.

Understanding geographical information in natural language or free text presents many challenges. Consider the query "beaches near Calgary". It consists of three important parts, a theme (*beaches*), a spatial relationship (*near*) and a location (*Calgary*). As is typical of most queries in information retrieval, it is under-specified and ambiguous. The geographical nature of the query delivers a number of additional challenges, many of which are difficult to address through standard information retrieval techniques. For example, it is unclear which Calgary is referred to (the landlocked but populous capital of Alberta in Canada, or the beautiful Calgary Bay, found on the Scottish island of Mull). Furthermore, what does "near" mean in such a context? Is it simply a set of beaches ranked by distance from some point location (the centre of downtown Calgary for instance), or all of the beaches found in some constrained space (e.g. all of the beaches on Mull), or beaches within some context-specific travel time or distance from Calgary?[2] Finally, beaches themselves can range in length from a few tens of meters to many kilometers, raising the question of appropriate ways of representing, ranking and comparing documents describing different beaches in the user interface.

[1] https://searchenginewatch.com/sew/study/2343577/google-local-searches-lead-50-of-mobile-users-to-visit-stores-study

[2] Note that however "near" is defined for beaches, it will have a different definition when the reference and user location change. "beaches near Calgary" implies a different definition of *near* than "coffee near the Hilton" or "airports near Laramie".

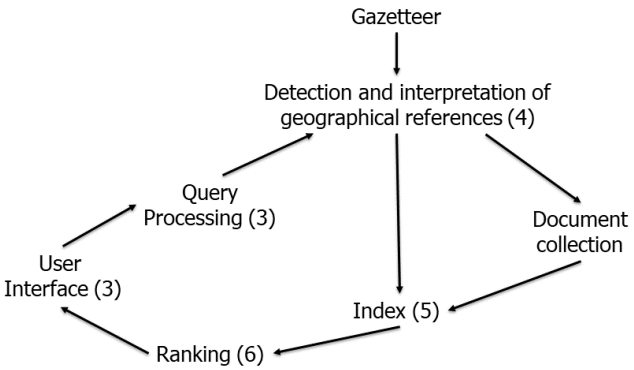


Figure 1.1: Schematic conceptual diagram of a GIR system and its related components - numbers indicate chapters where these topics are handled

Dealing with these, and numerous other challenges, lies at the core of what has been termed geographic[al] information retrieval[3]. The field was initially defined by Larson, 1996 as "an applied research area that combines aspects of DBMS research, User Interface Research, GIS research, and Information Retrieval research, ... concerned with indexing, searching, retrieving and browsing of geo-referenced information sources, and the design of systems to accomplish these tasks effectively and efficiently" (p. 81). Jones and Purves, 2008 refined this definition by emphasising, analogous to definitions of information retrieval, the importance of unstructured text: "GIR is therefore concerned with improving the quality of geographically specific information retrieval with a focus on access to unstructured documents such as those found on the Web" (p. 219). In the context of the example query given above, this definition has the important implication that GIR must be able to detect and resolve references to locations, typically but not exclusively in the form of place names, or more formally *toponyms*, from unstructured text documents.

Figure 1.1 illustrates schematically a basic model of a GIR system (from a more system-oriented perspective), which allows us to introduce the core concepts which will make up this article. The first component in such a system is a user interface, which mediates between the user and the system and supports user interaction, for example, helping users formulate queries, evaluate results and reformulate their search. With regard to query

[3]Further details about Information Retrieval systems can be found in IR textbooks, such as (Baeza-Yates and Ribeiro-Neto, 2011; Manning *et al.*, 2008; Hearst, 2009; White, 2016)

formulation it is important to consider cognitive representations of space and how these might influence natural language used to specify an information need that includes a spatial component. After a query has been formulated, it must be processed by the system to resolve under-specified information, or present the query and/or initial results to the user for reformulation. Typically, the query is then passed to one or more indexes and documents are retrieved, along with some information as to their (system defined) relevance. Based on these values, documents are ranked, and the results presented to the user in the interface. The representation of results may be optimised to allow browsing of large volumes of data or, in a more typical search paradigm, provide a ranked list of potentially relevant results. The user may then choose to refine the query, or click through to further information based on the results initially presented. Having designed a system capable of responding to queries with geographic content, an important research task is to demonstrate that a given system offers advantages over previously published work, through a thorough and reproducible evaluation.

Underlying such a system are a number of other key elements, always including a document collection itself, and typically a gazetteer (which records place names and associated information such as coordinates) or other structured geographical information. Document collections may simply be some part of the web, or more specific documents, for example relating to news stories, travel reports or mountaineering literature. The nature of these collections, for example in terms of their spatial distribution and biases or in terms of the target readership of a set of articles, have important implications for the design and evaluation of any resulting GIR system, and should be carefully considered. Gazetteers, which typically take the form of lists of place name related information, including geographic coordinates, form one of the key bridges between the disciplines of Information Retrieval (Baeza-Yates and Ribeiro-Neto, 2011) and Geographic Information Science (GIScience) (Goodchild, 2010). These disciplines, underpinned by research in computer science and geography, form two important areas from which much work in the field of GIR has emerged.

1.2 Exemplar GIR systems

As set out in the previous section, the focus of this article is on unstructured text, and methods which allow geographic references in such text to be identified and indexed, together with associated thematic information. By indexing such information it is then possible to both perform targeted searches using some form of query interface, and to explore content with respect to location and theme. In practice, as we will show in the following sections, much research on GIR has focussed on individual aspects of the process (e.g., on georeferencing, ranking, indexing or evaluation). However, in the last decade or so numerous authors have developed more or less complete process chains aimed at performing

Geographic Information Retrieval. These process chains vary widely: firstly, with respect to their purpose; secondly, in terms of the text corpora analysed; thirdly, with regard to the external resources used in, for example, structuring space and retrieving toponyms; and fourthly, through the different methods applied to carry out georeferencing, indexing, as well as querying and ranking. Here we introduce a number of exemplar systems and summarise key points with respect to the first three issues. The reader should note that the selection of systems chosen here is not exhaustive; rather these have been identified to illustrate different aspects of GIR in the remainder of this article. In selecting systems our focus was on published literature related to the systems, which have mostly been implemented in a research context and, as is typical in research, many of these systems are no longer maintained. Further examples and a comparison of systems can be found in Palacio *et al.*, 2010. In subsequent sections we will refer to the approaches taken by these systems where appropriate.

One of the earliest examples of a GIR system was GIPSY (Larson, 1996) which aimed to allow search within a so-called Digital Library. GIPSY focussed on analysing documents rich in geographic content, and used a gazetteer derived from the US Geological Survey's Geographic Names Information System (GNIS) for California. The Web-a-Where project was one of a number of early initiatives which linked locations from gazetteers to web pages (Amitay *et al.*, 2004). Web-a-Where used a number of corpora including a small (200 page) collection of .gov pages, mostly stemming from the US, and a second small collection from the Open Directory Project (ODP) with worldwide coverage. Gazetteer data were again derived from GNIS for the US, and from a number of other sources for non-US locations.

The SPIRIT search engine again focussed on web documents (Purves *et al.*, 2007), but used an initial corpus of some 94 million web pages to georeference 900,000 documents referring to locations in the UK, France, Germany and Switzerland. Gazetteer data were sourced from two datasets - firstly SABE (Seamless Administrative Boundaries of Europe) and, secondly, only relevant to the UK, the Ordnance Survey 1:50000 Scale Gazetteer. The STEWARD search engine (Lieberman *et al.*, 2007) was also initially developed to search general web documents, but also introduced the notion of searching on more specialised corpora, and in particular news articles.

News stories have proved very rich sources of corpora for GIR. Perhaps the most prominent example of an individual system is NewsStand (Teitler *et al.*, 2008), which focussed on collecting and effectively visually summarising news stories in real time, and required the use of gazetteers adapted to the geographic coverage of the sources used. NewsStand thus differed from the systems described previously in that the system was designed to deal with streamed, rather than static, content. Content sourced from newspapers and news wires formed the basis for many of the evaluation tasks in the cross-lingual geographic evaluation efforts known as GeoCLEF (Gey *et al.*, 2005; Mandl *et al.*, 2008a). Sources included, among others, the Los Angeles Times and the Glasgow Herald in English, Der Spiegel in German

and Público in Portuguese. The varying coverage of these collections, and their underlying languages, meant that systems had to adapt in terms of the supporting data used, with challenges emerging due to the local nature of coverage and resulting gaps in gazetteer based knowledge (Stokes *et al.*, 2008).

Methods from GIR have obvious applications in allowing corpora to be analysed and explored. One commonly cited class of corpora relates to cultural heritage. Thus, for example, the Virtual Itineraries in the Pyrenees (PIV) project focussed on extracting spatial and temporal information from a regional media library containing articles pertaining to the Pyrenees (Gaio *et al.*, 2008). As with NewsStand, this in turn implies that a locally adapted gazetteer is made available, so that more fine-grained toponyms can be identified. A similar approach, focussing on natural features, was taken by Derungs and Purves, 2014 to characterise spatial regions according to text used to describe them. This work analysed articles from the Swiss Alpine Club dating back to 1865, and used an administrative gazetteer provided by the Swiss national mapping agency to identify fine grained toponyms.

Both of the examples described above deal with corpora with primarily national or regional coverage. Other researchers have used GIR methods to summarise large textual corpora at coarser scales and for much larger geographic regions (ranging from the US to the entire world). The requirement for detailed information in gazetteers is correspondingly lower, and systems of this nature have also sought to use machine learning methods to georeference content without recourse to gazetteers. Key here is the availability of training data with coordinates, of which perhaps the most prominent example is GeoWiki - the set of Wikipedia pages associated with latitudes and longitudes. Two examples of such systems are FrankenPlace, which allows visualisation of thematic terms extracted from travel blogs and Wikipedia (Adams *et al.*, 2015) and an exploration of two much more specialised corpora, both of which nonetheless have global coverage, using the TextGrounder system (Brown *et al.*, 2012). A final example is recent work from Wang and Stewart (Wang and Stewart, 2015) who seek to link semantic information about hazards to both location and time.

1.3 Structure of the article

In this article we aim to bring together research on all elements of Geographic Information Retrieval, with a focus on methods which are applied directly to unstructured text[4], where geographically relevant information is present, but must be detected and annotated through the use of appropriate methods. Our aim in writing the article is to provide an overview of the research field, and in so doing identify key remaining research challenges in GIR.

In the following chapters, we start by exploring some basic geographic concepts, and reviewing ways in which space is represented and analysed in Geographic Information

[4]Our focus in this article is on written text, rather than social media.

Science. Chapter 3 then considers the issue of user needs in GIR and how these can be met by different forms of user interface, many of which are designed to assist the user in visualising the geographic context of the retrieved information. The next chapter, on georeferencing, addresses the fundamental problem, highlighted in this introductory chapter, of identifying references to geographic location in unstructured text, typically in the form of place names, and resolving these references to specific locations on the Earth's surface. The following chapter on indexing reviews various approaches of spatio-textual, or spatial keyword, indexing, that integrate techniques of text indexing and spatial indexing to provide efficient access to large document collections in ways that support user queries that refer both to thematic concepts and to geographical location. The challenge of how to rank the resulting retrieved documents to take account of both the thematic and geographic context is the subject of Chapter 6 on relevance and ranking. Key to improving the quality of retrieval systems in ways that help to make the results useful and accessible to the user is the effective development and implementation of schemes for evaluation, which form the subject of Chapter 7. Each of Chapters 3 to 7 refers to some of the exemplar systems that were introduced in the current chapter. This is either in a separate section of the chapter or, in the case of Chapter 3, they are referred to directly within parts of the chapter. We conclude by presenting a number of research challenges that we believe need to be overcome in order be successful in advancing the field of geographic information retrieval.

2 Basic Geographic Concepts

2.1 Introduction

Research in GIR is, as we set out in the introduction, fundamentally interdisciplinary. Our aim in this chapter is to introduce a set of basic concepts which underly many of the assumptions implicitly made in GIR with respect to ways in which space (and place) is conceptualised, represented and analysed. To do so we focus on three aspects: ways in which language is used with respect to space (Section 2.2), ways in which space is typically modelled (Section 2.3) and some simple operations which can be carried out on spatial data (Section 2.4). Our aim here is to provide the interested reader with a good starting point for following up these themes, and to make explicit some commonplace assumptions in GIR.

2.2 Spatial language in documents

In the previous chapter we identified the theme, spatial relationship and location as the three main components of free text queries in a GIR. In literature on spatial natural language (e.g., (Herskovits, 1986; Coventry and Garrod, 2004)) analogous terminology is used to describe

a spatial scene. Thus the theme, i.e. the objects or phenomena that we wish to locate, is referred to variously as the *located object*, the *figure*, or the *locatum*. This is the part of a scene that is the focus of the description. The location to which the *located object* is spatially related can also be referred to as the *reference object*, the *ground* or the *relatum* (Talmy, 1983; Landau and Jackendoff, 1993). For example, in the phrase "beaches near Calgary", "beaches" takes the role of located object (of which there could be several), "Calgary" is the reference object, and the spatial preposition "near" specifies that in this description the located objects are in the vicinity of the reference object. The phrase is typical of many such sentence structures in which the spatial relationship is usually indicated by a preposition (though verbs can also be used). Other common spatial prepositions or prepositional phrases in English include, for example, *in*, *at*, *on*, *north of* and *in front of*[1].

We can see from the above example that in practice in natural language descriptions, both the located object and the reference object can be given place names that can vary in geographical granularity from fine grained objects, such as a building to extensive regions such as London. Sometimes one or other of the located and reference objects may not have an explicit name, as in "the building in front of me". In Chapter 4 we review methods for automatically detecting the presence of terms that refer to locations, typically place names (or toponyms), and referencing them to specific coordinates that provide their position on the Earth's surface.

A limitation of much of the work in GIR to date on attaching geographic footprints to documents is that the georeferencing is often based solely on the individual place names, without attempting to understand the meaning of the locational expressions that include the place names. As GIR develops we can expect to see more emphasis placed upon interpreting complete natural language localisation expressions. In doing so we distinguish here between three types of frame of reference that affect the interpretation[2]. These are *relative*, *absolute* and *intrinsic*. In a relative frame of reference, the spatial relationship depends only on the relative locations of the located and reference objects, such as "the cathedral in York" or "castles near Bamberg". In a frame of reference that employs so-called projective relations, such as left, right and behind, the interpretation depends upon the orientation of the viewer (i.e., which way they are facing). In an absolute frame of reference the spatial relation relates to an external reference frame that is independent of the located and reference objects. The prime example of this is the set of cardinal directions that relate to the Earth's axis of rotation, as for example in "Camp sites north of Paris". An intrinsic frame of reference relates to the inherent properties of the reference object, as in "the clock tower in front of the palace", in which it is assumed that the palace has a front and a back that do not

[1] We confine our discussion here to English. Note, however, that spatial prepositions often do not map trivially between languages.

[2] The interested reader can find a comprehensive review of the notion of frames of reference from differing disciplinary perspectives in (Levinson, 2003a)

depend upon of the orientation of the building. A further issue that affects automated interpretation of locational expressions is the fact that many spatial relations, such as near and north are essentially vague with a range of possible interpretations in a given context; while some place names are also vague with boundaries that are fuzzy in their definition.

Finally, it is also important to note that differences occur in the ways in which space is described in language according to the form of communication (e.g., in direct speech or in a written document), the purpose (e.g., route directions versus a novel), language (for example some languages use exclusively absolute reference frames) and the device and formality of communication (e.g., descriptions of locations in social media content may be very different from those in newspapers). This in turn means that care should be taken in implying general rules with respect to the way in which language describes space, while at the same time appreciating the extent to which theories of spatial language can facilitate research.

2.3 Basic models of space

In the following we briefly introduce some key notions with respect to ways in which space is modelled in computers. This material is designed to serve as a brief introduction, and we refer the interested reader to one of a number of GIS texts as a starting point for exploring these topics in more depth such as (Longley *et al.*, 2015; O'Sullivan and Unwin, 2014). It is important to understand these spatial models in GIR, since whenever information in a document is referenced to geographic space it is these concepts that are used to represent geo-spatial location whether in quantitative or qualitive form.

Many aspects of information querying, analysis and visualisation in GIR depend upon methods of spatial data representation that were developed in the context of GIS. The main computational models of spatially referenced information fall into two categories: object or entity-based and field-based models (Figure 2.1). In object-based models, the location of individual features such as buildings, roads and rivers is represented explicitly through geometric primitives or objects. In field models the embedding space rather than the individual objects takes the primary role and each location within space can be assigned attributes relating to the objects or phenomena present at the location.

2.3.1 Object-based spatial models

In the object-based model the primary geometric primitives used to represent the location of two dimensional objects are points, lines and areas. The location of a point in 2D is represented by a single pair of coordinates. An area primitive, often also referred to as a polygon, is bounded by one or more lines, where a line is the locus of a sequence of points and is bounded at each end by a point. When a line is represented by more than just two

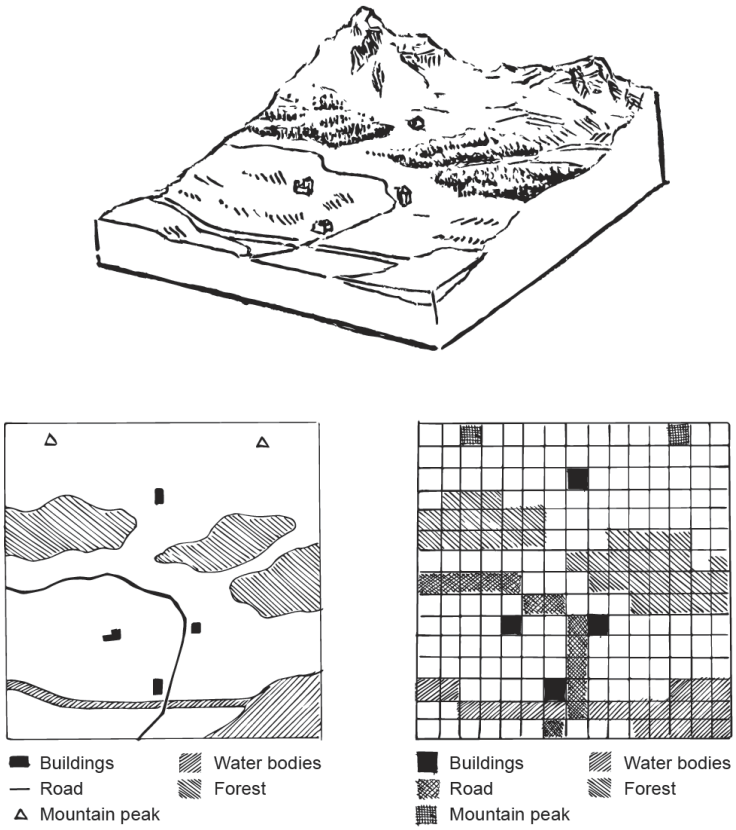


Figure 2.1: Abstracting from the real world to either an object-based or field-based model of space. Image adapted from an original by Pia Bereuter and licensed under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

points (i.e., the start and end) it can be described as either a polyline or a line string. While a line is typically represented by a sequence of point coordinates, it could alternatively or additionally be represented by a mathematical function, such as a spline function that is constrained to pass through or near to a given set of points. Note that an area in 2D can also be regarded as a type of surface, i.e. one that is planar.

Which geometric primitive is used to represent the location of a geographic feature typically depends upon the level of abstraction or, in cartographic terms, generalisation. Thus on a small scale (less detailed) digital map the location of a city could be represented geometrically by a single point that in a cartographic representation was associated with a point-referenced symbol, such as a circle or disc. In a slightly more detailed map the city could be represented as an area with a polygonal boundary. With further detail, a city could be represented by a complex combination of lines to represent road boundaries, or road centre lines, while polygon boundary geometry could represent buildings and parks, for example.

To assist in querying and analysing digital map features a further level of geometric abstraction is sometimes introduced and referred to as *topological structure*. In the context of GIS topology refers to the nature of connectivity between geometric objects and it is associated with several topological primitives that can be used to represent the relationships between (in 2D) point, line and area features. These primitives are a *node*, which corresponds either to an isolated point or the terminus of a line; an *arc* that corresponds to the path of a linear feature that either connects one node to another; a *ring* consisting of a closed loop of one or more arcs, corresponding to the boundary of an area; and a *face* that represents the area that is bounded by a ring. These topological primitives are typically used to represent either polygonal maps or network maps. Polygonal maps consist of a set of areas that share boundaries with each other (consider for example the states of the USA, or a set of land parcels). Network maps consist of a set of arcs, that connect with each other at nodes. The arcs could correspond to real world features such as roads while the nodes could correspond to road junctions. In practice, nodes, arcs and rings are usually associated with co-ordinates in the form of points, lines and polygons. Given a topological model it is straightforward to derive a set of topological relationships such as containment, adjacency and intersection between polygons and to test for properties such as planar enforcement (which implies that all of the space within a region is filled, and any one location may only lie in a single polygon).

2.3.2 Field-based spatial models

The field model of spatial data, in which each location in space is associated with some measured attribute or category, is most commonly represented by a raster data structure. A raster represents space as a 2D matrix of rectangular grid cells, each of which is associated

with a data value that characterises the region of space occupied by the grid cell. Alternative forms of data representation for a field include a single polynomial that can be evaluated at all locations in the region of interest, or a set of local polynomial "patches" that cover the given region. In the context of GIR, field models are less commonly used than object-based models. This is because geographic entities mentioned in text documents are often treated as discrete features, such as cities or buildings that are most easily represented with a geometric primitive, such as a single point or a polygon.

2.3.3 Crispness, vagueness and uncertainty

The geometric primitives of points, lines and areas associated with object-based spatial models are well suited to representing geographic features that are regarded as having a definite and relatively precise location in space.

Some features, such as buildings and roads, can indeed reasonably be treated as having such a crisp definition. Other features, such as regions of vegetation and soil of a particular type, rarely have a crisp boundary in the real world. It is also the case that many named places are essentially vague in their spatial extent in that there is no agreement on exactly where the boundary is located. Examples of such places are city centres, the "Mid West" in the USA and "the Midlands" in England (Fisher, 2000; Schockaert, 2011).

For such vague or fuzzy regions, field based models may be more appropriate as they enable a representation, such as a raster model, in which each cell can be associated with a confidence value. In these representations it is often the case that there is a central core region on which there is complete agreement that the respective name applies and a border zone in which confidence in membership of the region reduces with increasing distance from the core, depending upon personal perception and experience (Cohn and Gotts, 1996; Montello *et al.*, 2003).

While the object-based or vector model is typically associated with crisp representations of geographic phenomena it is very important to appreciate that all measurements of location in geographical information systems are subject to error and hence to uncertainty. This can arise variously from the original survey measurements, from representation of a real world extensive object as a point or a line, and from transformations of data that may take place in for example converting from one coordinate system to another or between data formats. Error in the recorded locations of spatial objects can lead in turn to error in the results of queries and analyses that compare or combine one object with another, such as when determining whether a point referenced object lies inside an area referenced object (where the point and boundary respectively could be inaccurately represented).

2.3.4 Coordinate systems

Geographic coordinate systems provide a quantitative method of recording location relative to some agreed reference location or datum, i.e. the origin of the coordinate system. Geographic coordinates in the form of latitude and longitude values record angles relative to reference planes. In the case of latitude, this is the angle between the plane of the equator of the Earth and a line between the centre of the Earth and the location being recorded on the Earth's surface. Longitude is the angle between a reference plane, that contains the Earth's axis and includes the Greenwich meridian, and another plane that also includes the Earth's axis but intersects the Earth's surface on another meridian line, i.e. the measured line of longitude that passes through the measured location.

When measuring latitude and longitude it is necessary to agree upon a geometric shape to approximate the surface of the Earth. This shape, which is described as a spheroid, is normally an ellipse of rotation, i.e. a planar ellipse the shorter axis of which coincides with the Earth's axis and which is rotated about that axis to sweep out a surface. An individual spheroid is defined by the length of its major and minor semi-axes. One such widely used reference spheroid, or *datum*, is called WGS84 and it is with respect to this datum that GPS (Global Positioning System) coordinates are measured.

While geographic coordinates of latitude and longitude provide a single global coordinate system they are not a very convenient form of coordinates for purposes of measuring distance and areas on the Earth's surface. Note, for instance, that the distance between two lines of longitude that are one degree apart varies between zero at the poles and about 111km at the equator. For purposes of convenient measurement and plotting of features on maps, latitude and longitude coordinates are projected (conceptually along rays originating at the Earth's centre) from their locations on a spheroid onto a localised planar surface touching the Earth's surface either at a point or along one or more arcs. A regular rectangular grid coordinate system (typically in meter units) is then constructed on the planar surface and forms the resulting map projection.

In an azimuthal projection the projection surface is a flat plane tangential to the Earth's surface. In a conical projection a plane is wrapped around the Earth in the form of a cone that touches the Earth along a line of latitude, or some other circle on the surface. In a cylindrical projection a planar surface is wrapped around the Earth either horizontally (touching the poles), vertically (touching the equator), or obliquely along some other great circle.

A very large number of different map projections have been used to map the Earth's surface. Most countries, and states within larger countries, choose a particular type of map projection to create maps for their region. Great Britain, for example, is mapped by its national mapping agency, the Ordnance Survey, on a cylindrical projection referred to as the British National Grid based on a datum called OSGB36. That map projection is closely

related to a widely used set of cylindrical map projections called Universal Transverse Mercator (UTM) in which the cylinder touches the Earth along a selected meridian (or pair of meridians) in the vicinity of the region to be mapped. UTM projections are categorised into standard 60 zones relating to 60 lines of longitude, at 6 degree intervals, on which the projection can be centred. For a detailed account of map projections see (Bugayevskiy and Snyder, 1995).

Measurement of distance

The distinction between geographical coordinates, based on latitude and longitude, and map projection coordinates, based on a rectangular grid in meter units, is a very important one when making measurements of distance. On a map projection distances can be measured easily and consistently using the normal formula for Euclidean distance. Unfortunately there is no single map projection on which this can be done for the whole Earth (since, as explained above, such map projections relate to local regions of the Earth). Latitude and longitude provide a global system of coordinates, but measurement of distance (and areas) requires the use of formulae based on trigonometric functions that take the angular coordinates as their parameters. An example of a simple method of measuring distance this way is the Haversine formula[3], which approximates the shape of the Earth as a sphere (i.e., with a single radius) and is a distance along a great circle on the sphere. The results are adequate for many applications of GIR, particularly for shorter distances, but more accurate measurement of distance with latitude and longitude requires geodetic methods that model the shape of the Earth as a spheroid (Karney, 2013; Rapp, 1993). A common method for computing distance on the spheroid uses the Vincenty formulae [4]. In practice the difference between distances measured when using a spherical as opposed to a spheroidal representation of the Earth is usually less than 0.5%[5].

2.4 Basic methods for handling spatial data

Geographical information systems provide a wide range of functionality for operating on geographical information. These can be categorised broadly into those concerned with preprocessing data to make it usable for some purpose; querying or retrieving information at a particular location; and analysing geographical information (Longley *et al.*, 2015; O'Sullivan and Unwin, 2014). Examples of preprocessing operations include those concerned with changing from one coordinate system to another, transforming between raster and

[3] http://www.movable-type.co.uk/scripts/gis-faq-5.1.html

[4] https://www.movable-type.co.uk/scripts/latlong-vincenty.html

[5] For a discussion and examples of these differences between using a spherical as opposed to a spheroidal representation of the Earth see https://www.r-bloggers.com/great-circle-distance-calculations-in-r/

vector spatial models, building topologically structured datasets, re–classifying data between one form of classification system and another, and reducing the level of detail or abstraction of geometric and semantic representations. Functionality for retrieving from a spatial database allows the user to formulate queries that specify locations quantitatively with coordinates, or in terms of spatial relationships to one or more named objects or themes. There is a considerable variety of methods for spatial analysis of geographical information, major types of which relate to measurements of distances, areas and volumes, spatial patterns in the distribution of phenomena, correlations between occurrences, routes within networks and across surfaces, interactions in space and time, including the flows of goods and people between locations, and the intervisibility of locations.

Here we confine ourselves to a summary of the main types of spatial query that are typically employed when retrieving information in a GIR system.

Spatial query methods

Spatial query methods are characterised by the use of three main types of spatial relationship to specify the relation between what is to be retrieved and some other reference object or location. These types are *proximity*, *topological* and *orientation*.

Proximity query In a proximity query (Figure 2.2), data are retrieved within some distance of a reference object. Examples in GIR would be to find documents that relate to locations that are within a specified distance of a given city boundary or within some distance of a river or road. Rather than specifying an exact distance, a user might specify a relation of *near* with no precise constraint on the distance (which would be context dependent and would present a challenge to automated interpretation). The locations of the documents would be represented by a document footprint typically in a form such as a point or a rectangle and it is this geometry that is compared with the location of the geometry of the reference object in order to measure distance.

Topological queries Topological relations define the nature of the connectivity between a pair of geometry objects. The most common of these relations are *inside* (conversely *contains*), in which the located object geometry is entirely within the extent of the reference geometry, *meets* or *touches* in which the boundary of one object coincides with the boundary of another, *overlaps* in which only a part of one object is coincident with the inside of the other, *equals* in which the two geometries are identical, and *disjoint* which refers to complete separation of the objects. These relations have a role in GIR queries particularly where the reference object in the query is represented by some "reference geometry" that might be obtained from a national mapping agency. A simple example would be to find documents whose geographic scope lie inside the boundary of a specified country. Another example is

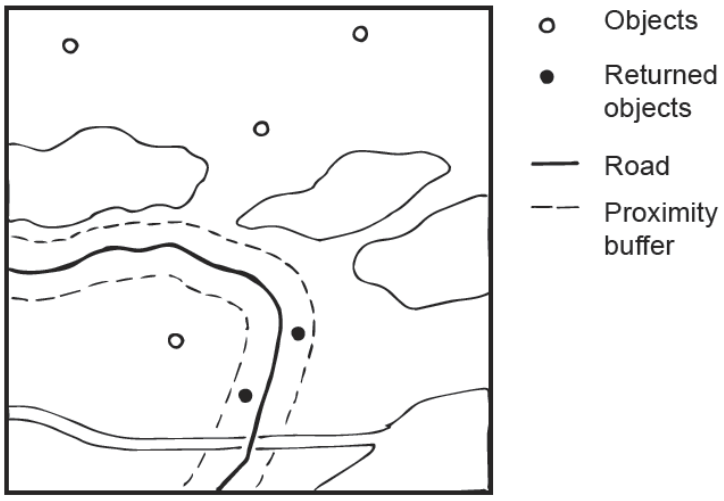


Figure 2.2: A proximity query using a buffer (a crisp region) specified here as some distance from a road. Image adapted from an original by Pia Bereuter and licensed under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

when retrieving documents that relate to the neighbouring countries of, say, Switzerland. Then the reference geo-data could be used to find the countries whose boundaries share a boundary with Switzerland (e.g., Figure 2.3), while the geometry representing the document footprint was used to test for containment within the respective neighbouring country boundaries.

Orientation or direction queries While direction is widely referred to as one of the main types of spatial relationship, it is rarely implemented in geographical information systems or spatial database query operators. An example query would be to find documents relating to areas "north of London", but because "north" is vague and context dependent there is no agreed standard interpretation. Equally it is not common when querying geo-spatial data to wish to specify directional relationships quantitatively, as a particular angle. Doing so would again raise the question of how exactly the angular relationship should be measured between different forms of geometry object (c.f. Figure 2.4).

2.5 Summary

This chapter has provided an introduction to some of the main geographical information systems concepts that are relevant to GIR. We started by reviewing some fundamental concepts that underly descriptions of geographic location, distinguishing between Located

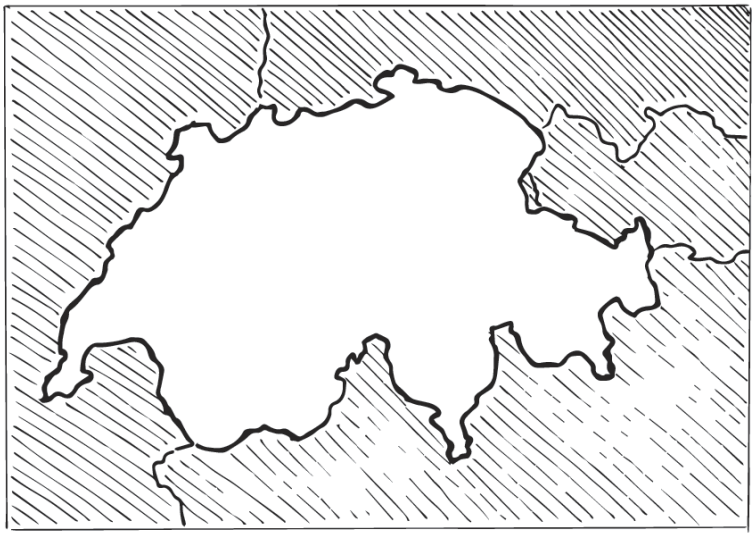


Figure 2.3: A topological query specified all the countries touching (adjacent to) Switzerland. Image adapted from an original by Pia Bereuter and licensed under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

Objects (LO), Reference Objects (RO) and their linking spatial relationship, and between several spatial frames of reference, notably *relative* (between LO and RO), *absolute* as in the cardinal directions, and *intrinsic* that relate to inherent properties of an object (such as its front). We distinguished between the two major types of spatial data model of object based, or vector, that represents entities with geometric primitives (points, lines and areas) and field-based or raster models that record what is present at a predetermined set of locations (typically a rectangular matrix of cells). A distinction was made between crisp, precisely defined entities and vague or uncertain phenomena that might reflect variation in human perception, or a lack of adequate data. Methods of recording and modelling people's perception of the locations of places remains an important research challenge, though several papers have been published on the subject (e.g., (Montello *et al.*, 2003; Gao *et al.*, 2017)). The importance of understanding the basics of coordinate systems was emphasised, as there are many different coordinate systems with a major difference between global systems, as in angles of latitude and longitude, and local systems that use meter units resulting from the projection of locations on a part of the Earth's curved surface to a planar grid. When querying geo-spatial information there are three common types of query that differ according to the major types of spatial relations. These are topological relations, that specify forms of connectivity or coincidence in space, such as inside, touch and overlap; proximal relations that refer to distance either quantitatively (as with meters) or qualitatively (as with words such as *near*); and orientation relations that specify direction quantitatively with angles or

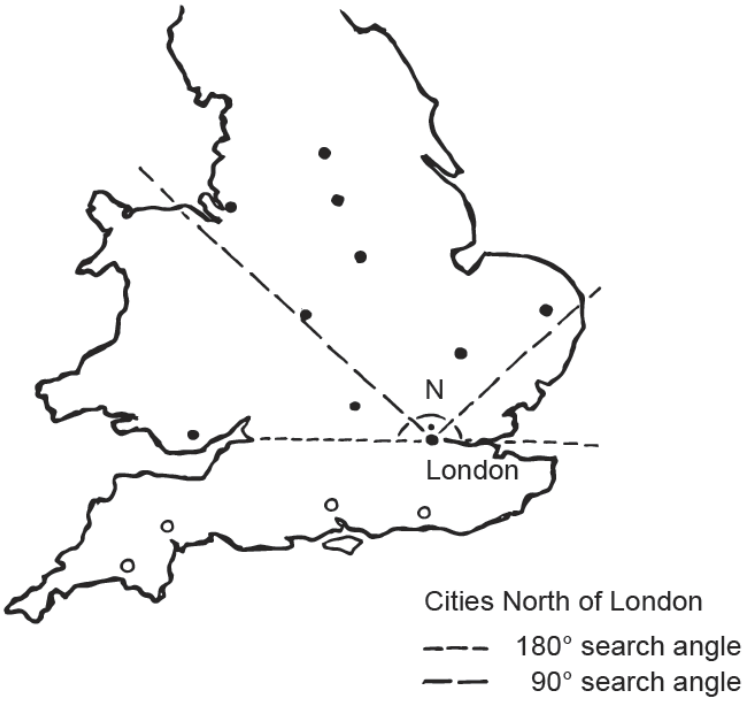


Figure 2.4: Specifying a directional query using two different angles for North, and treating the ground (London) as a point at a national scale. Image adapted from an original by Pia Bereuter and licensed under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

qualitatively with terms such as *north* and *in front*. It is notable that commercial GIS do not normally support interpretation of the qualitative relations of proximity and orientation as they are essentially vague and their successful automation remains as a challenge. With respect to research in GIR, we emphasise the importance of understanding ways in which other disciplines can provide theory and empirical inputs to research. For instance, systems which seek to implement methods with respect to vague spatial relationships such as near can usefully gain by understanding ways in which vagueness has been conceptualised (e.g., Fisher, 2000), explored in language and cognition (e.g., Levinson, 2003b), represented computationally (e.g., Cohn and Gotts, 1996) and analysed (e.g., Derungs and Purves, 2016).

3 User Needs and User Interaction

3.1 Introduction

In this chapter we introduce ideas from conceptual and empirical studies which can help us to understand typical information needs in GIR (Section 3.2). From the perspective of the user, these information needs must be captured by the user interfaces of the GIR systems that they employ. Exploring the nature of the user interface and its ability to deal with typical information needs from the perspective of GIR forms the second main element of this chapter (Section 3.3).

3.2 Information needs in GIR

People often use GIR systems to seek for information, which could arise from a person's incompleteness in knowledge or a particular problem to solve. This problem could be externally motivated (e.g., a work task or student assignment) or internally motivated (e.g., curiosity). People's use of GIR systems will be driven by their *information needs*[1]. Searching for information then aims to find relevant information in an attempt to address these needs. This could include more analytical approaches to finding information through the use of keyword searching, or more undirected forms of browsing, perhaps through the use of subject categories or visualisations.

Understanding people's information needs is a central but often neglected challenge in IR generally and GIR more specifically. In the past, it was often assumed that information needs can be relatively straightforwardly formulated as queries, and that this can be achieved by those developing systems with limited reference to the humans who actually use such systems. However, in practice this has led to claims with respect to information needs which may not be reflected in actual search behaviour, and in turn, a lack of attention being paid to potential information needs which have not been modelled. Information needs are closely coupled to notions of relevance and ranking, since an effective ranking scheme should rank more highly those documents which more closely match an information need (c.f. Chapter 6).

Here we take two perspectives when exploring information needs with respect to GIR. Firstly, we examine a conceptual model developed originally as an aid to analysing and classifying picture subjects, but widely applied in image retrieval more generally. Secondly, we review the nature of geographic information needs based on empirical data derived from analysing query logs.

[1]The notion of an information need highlights that a gap in knowledge exists between what someone knows and what they need to know to fulfil some kind of underlying goal, problem or work/leisure task (Cole, 2011).

The Panofsky-Shatford facet matrix was introduced in Sara Shatford's seminal paper (Shatford, 1986). It is based on theoretical notions with respect to the meaning of images with the aim of creating a way of organising and classifying image content. Importantly, Shatford concerns herself not only with literal meaning (is this a picture of Ross Purves or simply of a man?) but also the metaphorical or allegorical meaning that might be conveyed by a picture (is this a happy man?). Shatford thus proposes three categories: the *specific* (Ross Purves), the *specific generic* (a man) and the *about specific* (a happy man). Having identified these categories, Shatford then proposes four useful and simple facets for her classification based on questions which we might typically ask of an image: *Who?*, *What?*, *Where?*, and *When?*. These questions, along with Why? and How?, are argued to be key to a well-written news article by Teitler *et al.*, 2008. They emphasise the importance of Where?, and make the claim that questions dealing with either features (e.g., 'where did story X happen?') or locations ('what is happening in location Y?') are central to aggregating and retrieving news articles effectively.

We argue here that the 'where' facet from the Panofsky-Shatford facet matrix provides useful and important clues about potential information needs in GIR, which, as demonstrated by Teitler *et al.*, 2008 can be expanded beyond images. Firstly, the *specific of where* facet effectively refers to instances of locations which can be grounded, for example through a toponym (e.g., New York, London). Secondly, the *generic of where* facet implies that images may represent a location class, without necessarily being named (e.g., a town, bridge or village). Finally, the *about where* facet refers to ways in which places may be symbolized in more abstract terms (e.g., does a picture of this place represent notions of paradise for some users?). The Panofsky-Shatford matrix has proved very powerful in exploring ways in which queries can be categorised (e.g., (Armitage and Enser, 1997; Purves *et al.*, 2008)) and defined and could provide a useful means of exploring information need in GIR.

However, query log analysis in IR more generally has developed categories based around the nature of queries submitted to real operational systems, and thus also reflecting, at least partially, the expectations and nature of web content over time, with limited attention paid to specifically geographic aspects related to queries. Prominent amongst early studies of query logs was the work of Spink, Jansen and colleagues. They studied a wide range of query properties, but of most interest here are attempts at query classification. In their 2001 study (Spink *et al.*, 2001) the categories have little or no explicit link to location, other than in the class 'People, places, things' though many may implicitly relate to locations (e.g., culture, society, government and travel will often relate to specific local information needs rather than more generic questions about what is society or government). Notable in their work on temporal trends, despite the use of very small samples, was the increasing importance, in the period between 1997 and 2002, of queries related to 'people, places and things' (Spink *et al.*, 2002).

To our knowledge the first study which explicitly explored query logs from a geographic perspective was the work of Sanderson and Kohler who explored geographic questions in the query logs previously analysed by Spink (Sanderson and Kohler, 2004). Their study, though based on a very small sample, confirmed that a significant proportion of queries contained either a geographic term (18.6%) or more specifically a toponym (14.8%). Furthermore, they noted, unsurprisingly, that queries containing a geographical term were typically longer than average web queries.

A more comprehensive study of information needs and query logs was undertaken by Henrich and Luedecke, 2007 using the controversial search logs released, and then withdrawn, by AOL[2]. They explore information needs which were both implicit and explicit in terms of geographical constraints and identify through inference four main classes of geographical information need: (i) finding accommodation (e.g., hotels in a city or particular hotels); (ii) searching for habitation (e.g., homes for sale in a specific area); (iii) general information finding for a region (e.g., weather in London); and (iv) finding information about activities to do in leisure time (e.g., restaurants or things to do in a specific place).

More recent work analysing query logs has typically been performed, at least partially, in-house, and most researchers in GIR have little or no opportunity for access to such query logs. Xiao *et al.*, 2010 explored query logs in map search and made some important, and generalisable conclusions for GIR. Firstly, they noted that where a map-based interface was present, 90% of queries contained locations, a much greater proportion than that reported in previous work (c.f. (Sanderson and Kohler, 2004)). Secondly, users typically viewed fewer pages, which the authors speculate may be due to the specific nature of the results displayed (e.g., business names and addresses) making relevance judgements easier than in general web search. Finally, they observe that query locations correlate with population, and furthermore that users often query in such systems for nearby locations (49.2% of queries are in the same state as the user and 33% of queries are within 50 km of the user).

Such analysis and implied information needs suggest some key requirements for user interfaces supporting GIR. These include allowing the user to specify the spatial element of their query, typically with a map representation to support query formulation, results presentation and browsing as well as providing a means of filtering search results to a specific area. In the following section we introduce search interfaces in general, before focusing on key aspects related to GIR.

[2]These query logs became controversial since they were nominally anonymised. However, journalists from the New York Times quickly succeeded in identifying individual users, with queries including addresses being a key way in which users were found (http://www.nytimes.com/2006/08/09/technology/09aol.html).

3.3 User interfaces

The user interface handles the interaction between the user and the IR system and provides the means by which information is transferred. Among other things, the interface provides a way for the user to define their query, review the results of the query and modify their query (Hearst, 2009; Wilson, 2011; Russell-Rose and Tate, 2013). Common approaches that people use to find information include *searching* for specific items or information about a topic, *browsing* or using a combination of both approaches (Morville and Rosenfeld, 2006). The design and implementation of various search aids, together with good interface design, can help make search easier for end users and support them during their information seeking activities, whether by searching or browsing.

With respect to searching, Shneiderman *et al.*, 1998 describe a four-phase framework that summarises the main activities of information searching that must be supported by the interface: (i) formulation, (ii) action, (iii) review of results, and (iv) refinement. A user typically comes to the IR interface with an information need, which defines what information they are looking for. The user must express their information need using the functionality provided by the system, such as a search box (formulation). This is followed by some mechanism that enables the user to execute the search (action). The user must then evaluate the results to determine whether or not the retrieved documents satisfy the query, and more broadly their information need (review results). Often users will carry out successive searches, reformulating and refining their searches as they go (refinement). Most interfaces will also provide functionality to support browsing and exploration of the document collection, for example through the provision of controlled vocabularies, visualisations and hyperlinks (Marchionini, 2006). Table 3.1 provides a summary of features commonly provided to support users during search activities.

Since typical GIR interfaces are based around commonly used search metaphors for search, we briefly explore here some common aspects of IR interfaces. Modern search interfaces come equipped with a large number of interface components to support users during the search process. For example, the interface may provide a search box, results list, pagination controls for paging through the results, facet elements that support the user in narrowing down and filtering their search results, auto-complete and dynamic term suggestion features and related searches. Wilson, 2011, Hearst, 2009 and Russell-Rose and Tate, 2013 provide detailed discussions of the full variety of interface elements; however, the focus here will be on the elements used to help the user with formulating and refining their query, and the elements used to display the results.

Usually the first element the user interacts with is a text box in which they enter their query. In practice most search boxes follow a simple design consisting of a text input box, in which the user types the query, and the "search" button used to execute the search. The IR system processes the text in the user's query, often allowing the user to qualify their

Table 3.1: Example support functionality offered by IR systems (Sutcliffe and Ennis, 1998).

Area of support	Example UI functionalities
Query formulation	Specialised query syntax (e.g. Boolean operators)
	Natural language query input
	Advanced search editors
	Dynamic term suggestion / auto-complete
	Query-by-example
	Query-by-pointing (e.g. on concept maps or images)
	Pre-formed queries (e.g. from existing thesauri)
	Re-useable queries (e.g. from past searches)
Document selection and examination	Provide summary of results, e.g. number of hits
	View content of selected document
	View and mark/re-use results (relevance feedback)
Query reformulation	Automated term suggestion (e.g. related queries)
	Spelling error correction
	"More like this" functions
	Search within results
Browsing	Controlled vocabularies and classification schemes (e.g. faceted classification)
	Concept maps
	Navigational aids (e.g. hypertext)
	2D/3D visualisations

query with operators such as "OR" and "-" (meaning exclude) and the use of quotes to indicate an exact phrase match ("the phrase to search for"). An alternative approach is to provide the user with multiple search input boxes – an advanced search interface. In this style of input the user can specify facets of their query separately. This enables the user to combine searches for specific aspects of the data, for example searching by author, date range, keyword and type of document, each of which could be combined by operators such as "AND" or "OR".

With the rise of search on mobile devices, where typing a query can be difficult, voice-driven search interfaces have become more popular (Feng *et al.*, 2011). While in some cases these enable interaction between the searcher and the search system, the interaction is frequently restricted to the system asking for clarification on the search terms. Apart from that functionality, these interfaces differ little from the standard full-text search box, except that the query is now spoken instead of typed.

After the user has submitted the query, the interface displays the Search Engine Results Page (SERP, (Wilson, 2011), p.50)) to present resulting documents to the user. The SERP itself typically consists of a list of documents that match the query the user provided. Each document in the SERP is usually represented using its title and a brief summary of the document's content, known as a snippet. The snippet is chosen so that it ideally includes

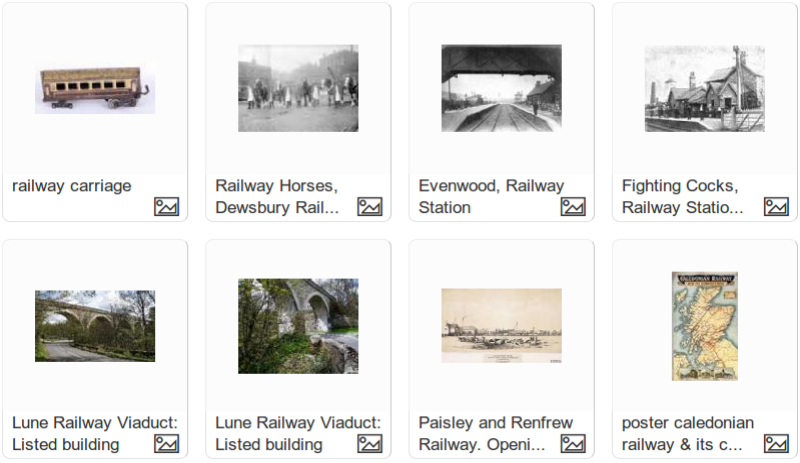


Figure 3.1: Alternative results presentation in the form of grid layout (Europeana, 2015)

those parts of the document that match the user's query, and thus relate to the information need as expressed in the query. This allows the user to judge whether the document is potentially relevant, without having to view the whole document. For primarily text-based results, this format dominates, but where the results are of a more visual nature, such as images or photographs of real-world objects, grid-based layouts (Figure 3.1) are often used, as they allow for a more compact display of more results, while also focusing on core aspect of the results, namely their visual impact.

3.4 Interfaces for GIR systems

The GIR interface extends the standard IR interface with additional functionality that enables users to specify the spatial constraints of their query and to view the spatial layout of results. In earlier GIR interfaces the user was presented with an additional text input box with which the user could specify the spatial constraint (Figure 3.2). In most interfaces the user would enter a toponym and the GIR system would, having disambiguated the toponym if necessary, handle this spatial constraint as part of the query sent to the search index (see, Chapter 5). For example, in the SPIRIT system (Purves *et al.*, 2007), the user could provide spatial constraints, such as "north of" or "near", to further restrict the spatial constraint and provide more flexibility than the basic "in toponym" search.

The main limitation of this approach is that it requires the user to know an appropriate toponym that matches the spatial area they wish to query, that is to say to have knowledge of the *specific of/ where* facet described above. If the user wants to search only part of the area defined by a toponym, has limited knowledge of appropriate toponyms, or wants to

Figure 3.2: Example local search interface (Google Local, 2005)

search within a region not easily specified by a toponym, then such an approach is limited. To address this, many spatial search interfaces also use a map representation that allows the user to specify a spatial constraint implicitly, for example by zooming or panning the map to specify a map region.

A further way of refining spatial queries is directly derived from classical GIS query mechanisms, and allows users to more precisely express a spatial query (Hill *et al.*, 2000). Here, a GIR interface allows the user to either specify a bounding box or a polygon as a query region directly on the map (Figure 3.3), which is then used by the IR system to spatially constrain the results. While this provides the user with more flexibility, it also introduces additional complexity, as the user now has to spend some time drawing their search area, and the actual retrieval process can be slower, since querying for documents in polygons is typically slower than simple bounding boxes. However, for particular application domains, such as real estate search, polygon-based search might meet user needs better than simple bounding boxes.

Many systems combine both approaches. For example, Figure 3.4 shows the interface for the SPIRIT system that incorporates both a text- and map-based query interface (Purves *et al.*, 2007). The text-based interface supports text input in a structured form, a triplet of `<theme><relationship><location>`, to define a query. The graphical interface, through the use of a map backdrop and basic interactivity, provides a user unfamiliar with an area with a means of specifying a query. For example, a user who is not familiar with local place names can draw a polygon approximating to a region of interest.

Early GIR systems often displayed results in a standard list-based SERP format. However, this provides no indication of the spatial distribution of the results and therefore map-based visualisations are now more widely used (see examples in Figures 3.5 and 3.6). Many map-based presentations simply display each result as a marker on the map, which the user can click on to view the result title and snippet. As the map only shows the spatial distribution of the results and not their relevance to the user's query, map-based SERPs often also include a standard vertical result list that shows the same results, but ordered by their relevance to the user (Figure 3.5).

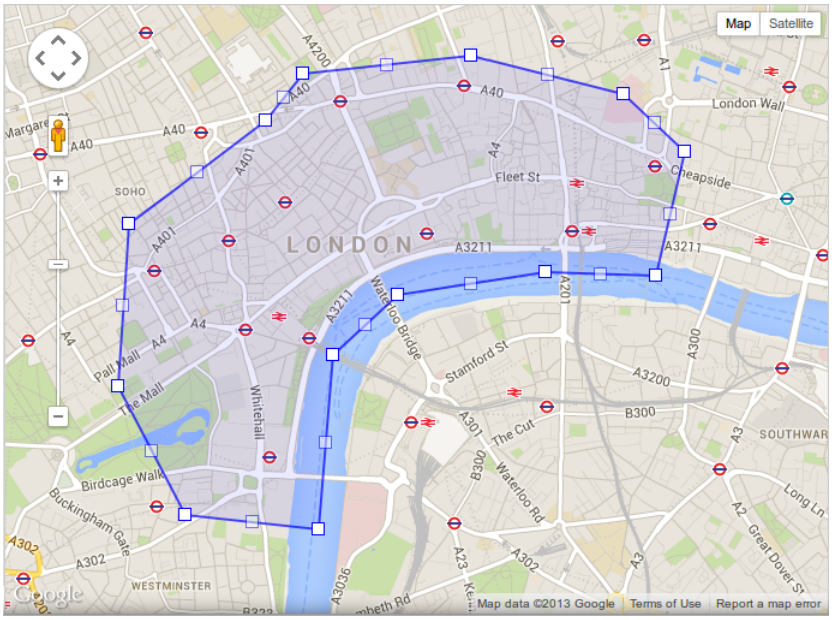


Figure 3.3: An example 'draw-a-polygon to search' interface (Rightmove, 2015)

The use of markers on the map restricts how many results can be shown; too many results lead to the markers overlapping, making the map unusable. Where the UI includes a standard search results list, the map usually shows the same set of documents that are shown in the results list. The user can then use the pagination controls to move through the list of results and the map is updated to show the markers for the currently displayed results. This allows for the display of large numbers of results without making the map unusable.

An alternative approach to this, that does not require a result list, is to cluster markers that are visually close on the map (Figure 3.6). Markers that are visually close together are merged into a single marker, which displays the number of results that have been clustered. The user can zoom into their area of interest and the cluster marker is then split into its individual markers. The advantage that this approach has over the inclusion of a result list is that it gives the user an indication of how much information they can find in a specific area. However, it also requires the user to interact and zoom into the map before they can view the result details, which slows down the user's interaction with the system. Such operations all belong to the general research area known as map generalisation (Mackaness *et al.*, 2011).

Figure 3.7 shows an example visualisation (for the query 'castles') from the interactive map interface of the FrankenPlace system (Adams *et al.*, 2015). Unlike other visualisations FrankenPlace makes use of a two-dimensional kernel density function with a bandwidth

Figure 3.4: SPIRIT system query and search results interface (Purves *et al.*, 2007)

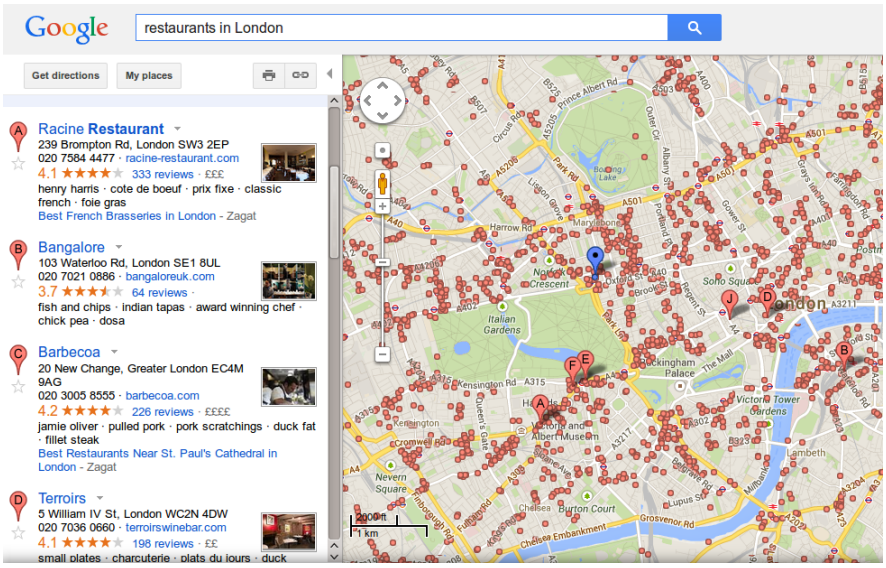


Figure 3.5: Interface showing the combination of map markers and a result list (Google Maps, 2015)

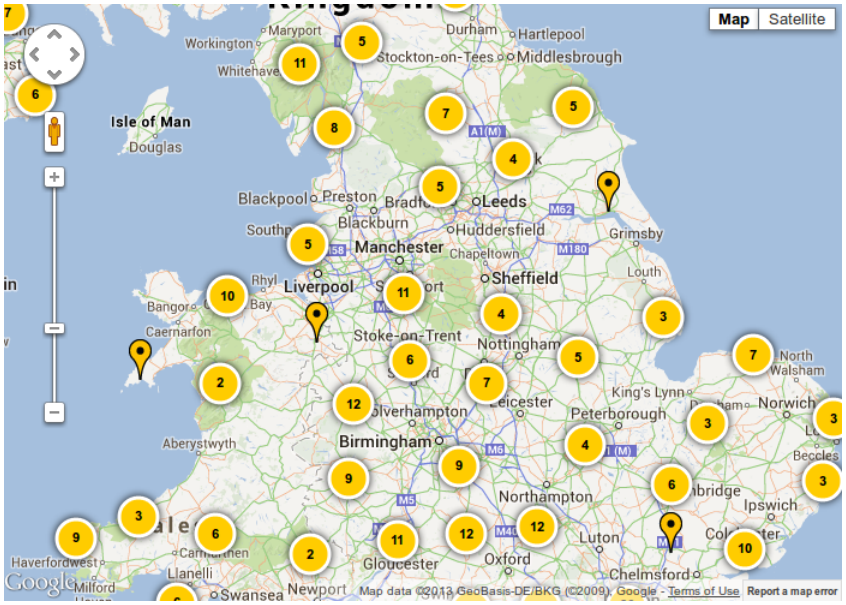


Figure 3.6: Example map-based interface that demonstrates the clustering of points (National Trust, 2015)

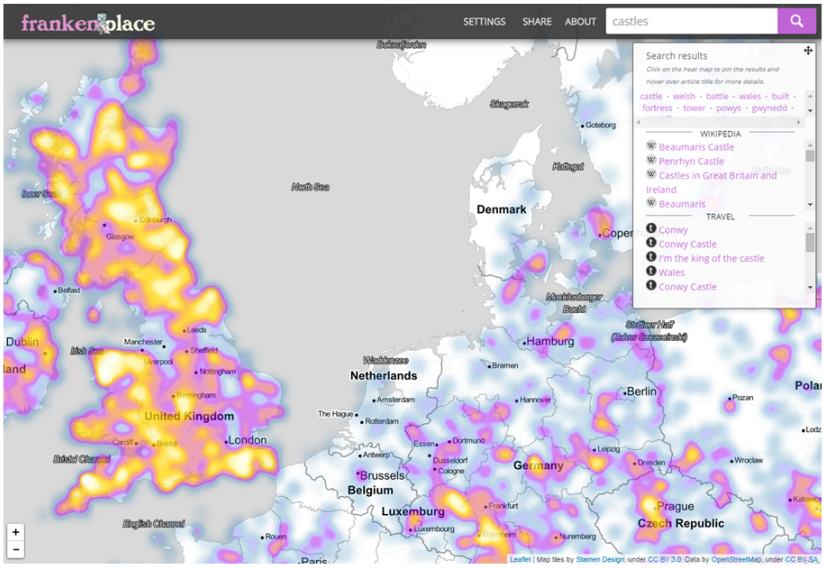


Figure 3.7: Example map-based visualisation from FrankenPlace for the query 'castles' (http://frankenplace.com/)

based on the zoom level to generate a smooth surface from locations mentioned in the search results. This produces a density surface and gives a useful initial overview of the results. The text box on the top right of the interface is generated dynamically and provides links to topics generated from the document collection during indexing and links to relevant Wikipedia articles for concepts from the user's query, zoom level and their current mouse position (in the case of Figure 3.7 the mouse is located over North Wales).

One of the limitations of current GIR interfaces is that they reduce the spatial relevance of the documents to a simple boolean choice: display the result or not display it. The detailed information about how well the document matches the spatial constraint is lost. Hobona *et al.*, 2005 demonstrate an alternative interface that enables the user to see how well the documents match both the semantic and spatial constraints. They describe a display in which documents are placed in a three-dimensional space depending on their semantic, spatial, and temporal relevance. This enables the user to quickly see how the documents relate to the query and also discover clusters of documents that share semantic, spatial, or temporal characteristics. However, such abstract spaces appear to be challenging for users to understand, and have gained little traction in GIR.

3.5 Recommendations for designing GIR interfaces

There are at least three aspects to consider when designing user interfaces for IR and GIR systems: consideration of the nature of user needs based on both conceptual models and analysis of query logs; the appropriate handling of user-system interaction; and creating an interface that is appealing and usable. An important mantra for user interface design is that people are better at recognising things they have previously experienced than recalling those things from memory. The use of menu-driven over command line interfaces is an example of implementing this theory of memory.

Existing guidelines should be used, such as the framework proposed by Shneiderman *et al.*, 1998 for developing interfaces to support the search process, the design recommendations for IR systems by Ahmed *et al.*, 2006 and general guidelines for usability (Nielsen, 1993). The classic 'golden rules' of interface design by Shneiderman *et al.*, 1998 are as applicable to the design of GIR user interfaces as standard IR ones. In addition, Nielsen, 1993 provides five attributes of usability that can be used to assess how usable an interface or system is: learnability, memorability, efficiency, errors (accuracy) and subjective satisfaction. This is important as studies continue to show that users still find IR systems difficult to learn, use, and remember (Ahmed *et al.*, 2006).

Vaughan and Resnick, 2006 describe a set of best practices developed to assist specifically in the design of search user interfaces. These design principles are organised into five domains: the corpus, search algorithms, user and task context, the search interface and mobility. Best practices include the use of faceted metadata within a controlled corpus, the use of spell-checking during user input, hybrid navigational support through combined search and browse, the use of past queries to frame the search context, the provision of a large query box, the organisation of a large set of search results into categories, showing the keywords in context in search results and designing alternative versions of content specifically for mobile and handheld devices. These are equally applicable to the design of GIR interfaces.

However, there are other aspects which are of particular relevance to the design of GIR interfaces, but which are, in most research to date, commonly ignored. Thus, although some of our exemplar systems (e.g., (Purves *et al.*, 2007; Teitler *et al.*, 2008; Brown *et al.*, 2012; Adams *et al.*, 2015; Wang and Stewart, 2015)) use map-based representations, which go beyond simply displaying individual documents as points, but seek to provide an initial overview (Shneiderman, 1996), they have not, at least obviously, been influenced directly by work emerging from either geovisualization (Dykes *et al.*, 2005) or cartographic theories of representation (MacEachren, 1995). It is also clear that little progress has been made to date in automated interpretation of spatial language in textual queries that use spatial relationships. This is an important challenge requiring effective models of the semantics of vague and context-dependent terms such as *near*, *north of* and *in front of*.

3.6 Summary

This chapter has considered the information needs of users and the subsequent design of user interfaces for GIR systems that help users meet their needs. The GIR system must provide functionalities that enable users to express their information needs (including the spatial element) in some way (e.g., as queries expressed in natural language or through the drawing of boundaries on a map). The interface must also support the refinement of users' queries and present results in an effective and usable way. We provided various example interfaces, including our exemplar systems, to illustrate example interface designs, arguing at the end the need to integrate theories from cartography and geovisualization into future GIR interfaces. In addition, emerging trends in designing interfaces to support the broader work tasks of users may also provide an important avenue of research in the future, particularly to meet the needs of professional users. Improvements to interface design may also be required to assist users with more exploratory forms of interaction (e.g., when their goals are unclear) and to better support users with limited (geographical) knowledge and ability to express their information needs verbally.

4 Georeferencing

4.1 Introduction

In this chapter we describe methods for identifying references to location as expressed in natural language text (Sections 4.2 and 4.3) and relating them to positions on the earth (Section 4.4). This is an important but challenging task due to the richness and ambiguity of natural language but a vital step in developing GIR systems. We also discuss related topics such as methods for automated assignment of geographical scopes to a document and the use of machine learning methods in georeferencing. We end by reviewing approaches taken in our exemplar systems (Section 4.8).

4.2 Georeferences

Within a text, references to locations are commonly referred to as *georeferences*. Georeferencing is a commonplace task in GIS, and involves associating information with some location in physical space. Georeferences can take a number of forms, for example all of the following are georeferences which may allow us to locate the beach at Calgary Bay in Scotland:

- Calgary, Mull - Postal address of the village

- Calgary Bay - Place name reference for the beach (also called a *toponym*)
- PA75 6QQ - UK postcode associated with 6 properties located near Calgary Bay
- 56.578397,-6.28006 - Latitude and longitude of Calgary Bay in WGS84
- NM 37250 51141 - Coordinates of Calgary Bay in the Ordnance Survey Great Britain projection, precise to 1m

Georeferences have a number of important and desirable properties (Hill, 2006). They should be unambiguous - that is they should refer to a single location only, which typically is the case within some given frame of reference. Nonetheless, there is no guarantee that there is only one Mull, or indeed only one Calgary on Mull. Georeferences should also be shared amongst those using them - thus, for example, if a letter was to be addressed using Calgary's Gaelic name *Cala ghearraidh*, it would be necessary for the UK postal service to be familiar with this form of the name. Georeferences should also be, insofar as possible, persistent through time. In practice, this is rarely the case (indeed in our example above the Anglicised version of the name probably dates from the first survey of the area), and even metric coordinate systems are subject to change (e.g., WGS84 was introduced in 1984). Finally, georeferences are typically associated with an implicit granularity - in our example the postal address of a sparsely populated village clearly refers to a coarser granularity than point coordinates given with a precision of 1m.

In GIR, our starting point is typically a document containing natural language text. Figure 4.1 shows a Wikipedia page that includes various forms of georeference, such as longitude/latitude coordinates, toponyms (e.g., Sheffield), and various other references to geographical regions (e.g., countries). When we read such a document, identifying references to location helps understand its geographical context. In doing so, we can easily identify important locations referred to in a document, and discard references to locations which are not relevant to the core theme. Performing this task automatically is a central, but complex, challenge for GIR, and much more challenging than the geocoding task carried out on more structured information typically used in GIS (e.g., adding a set of coordinates to a list of well-formed addresses (Zandbergen, 2008)). Although references to location in natural language can come in many forms ranging from *deixis* (e.g., "I am here") through the use of toponyms (e.g., Zurich) to postal addresses (e.g., Friesenbergstrasse 7, Zurich, Switzerland) and metric georeferences (e.g., "$47°22'N\ 8°33'E$") most work in GIR has focussed on dealing with toponyms and addresses (Leidner, 2006).

Given a document, our task is thus to both identify these geographical references unambiguously and, typically, assign spatial coordinates to them. Only by so doing, can we ensure that a geographical reference is associated with a unique location. Many different terminologies have been used to refer to this process – here we will refer to the process in

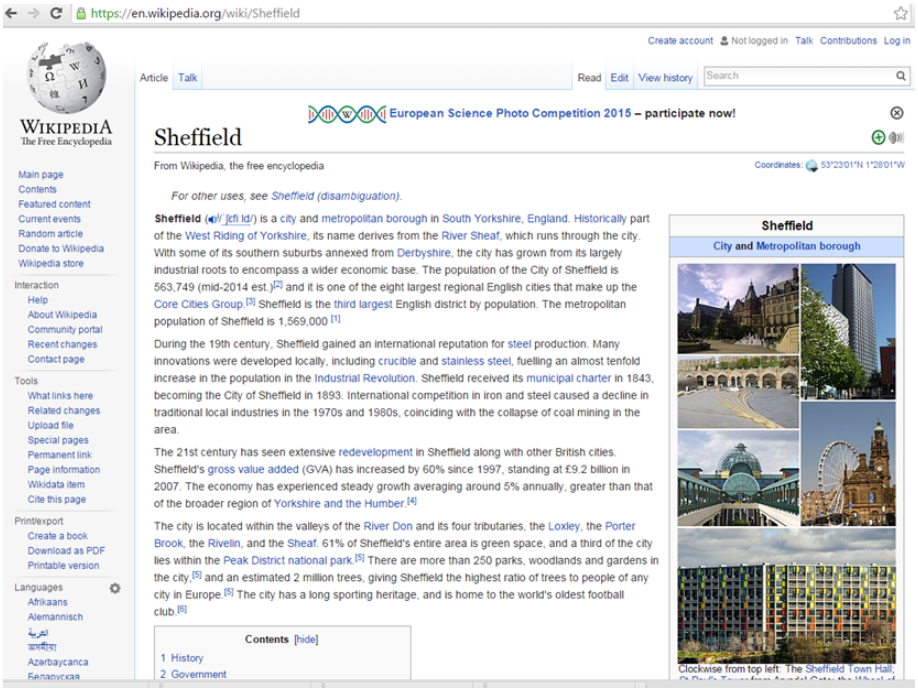


Figure 4.1: Example Wikipedia page with different forms of georeferences, such as names of towns and regions and latitude and longitude coordinates (Wikipedia, 2016).

terms of *geoparsing* (Section 4.3) and *geocoding*[1] (Section 4.4) following the definitions from Larson, 1996 and McCurley, 2001. After associating a reference to a location with a unique identifier, this identifier can in turn be associated with a metric georeference which typically takes the form of a point, line or polygon, though at least in principle such a georeference may also take the form of a raster cell. These georeferences may then be used to imply *document scope* – that is to say the location or locations with which the content of the document, or some part thereof, is assumed to be associated (Ding *et al.*, 2000; Andogah *et al.*, 2012). Computing document scope is discussed further in Section 4.5. In this article we focus on connecting linguistic references to locations (e.g., toponyms) to corresponding formal references of the locations (e.g., longitude and latitude coordinates), as well as on toponym identification and disambiguation.

[1] In GIS-based applications the notion of geocoding often refers to the process of converting addresses into geographic coordinates.

4.3 Geoparsing: identifying georeferences

A basic task in georeferencing is to identify candidate georeferences. For toponyms, this is commonly referred to as *geoparsing* or *toponym recognition* and can be likened to the Information Extraction (IE) task of Named Entity Recognition[2] (NER): the process of assigning every word or group of words to a set of pre-defined categories or entities, including "not an entity" (Nadeau and Sekine, 2007). These categories commonly include names, such as location, person or organisation, and numeric expressions, such as time, date and monetary amounts. For example, the text in Table 4.1 shows the typical output from a NER system[3] (where org=organisation and loc=location categories). Most NER systems involve a pipeline of stages, such as splitting a text into words and phrases (referred to as *tokenisation*), splitting documents into sentences, assigning Part-of-Speech (POS) information, and performing gazetteer lookup.

Table 4.1: Example output from identifying named entities.

Input: French company Aeroports de Paris International is the design consultant for the work. The company also operates Charles de Gaulle airport in Paris, where part of a terminal building's roof collapsed in May, killing four people.
Output: French company `<org>`Aeroports de Paris International`</org>` is the design consultant for the work. The company also operates `<loc>`Charles de Gaulle airport`</loc>` in `<loc>`Paris`</loc>`, where part of a terminal building's roof collapsed in May, killing four people.

Much previous work in NER has been to extract fairly coarse-grained entities, although more detailed location classes have been explored (Uryupina, 2003; Axelrod, 2003). The identification of place names may also involve assigning information indicating the type of place, e.g., city or town. In the case of identifying toponyms *semantic ambiguity* must be handled and resolved – in other words does "bath" refer to a place to bathe, or the city in the south-west of the United Kingdom. This problem, where the same name can be used for multiple categories is also commonly referred to as *referent class ambiguity* or *geo/non-geo ambiguity* (Amitay *et al.*, 2004).

Various approaches exist for identifying potential named entities in text (see Section 4.3.1). Regardless of the approach used, most NER algorithms combine lists of known locations (also called *gazetteers* (Hill, 2000)), organisations and people with rules or machine learning techniques that exploit elements of the surrounding context. It is worth noting that building toponym gazetteers which are both comprehensive (i.e., contain all relevant toponyms mentioned in a document) and accurate (e.g., do not contain duplicates of the same toponym, with slight variations in spelling or coordinates) is a challenging task (Sehgal

[2]This is also sometimes referred to as Named Entity Recognition and Classification, or NERC.

[3]This example is produced using the Stanford NER system: http://nlp.stanford.edu/software/CRF-NER.shtml

et al., 2006; Martins, 2011; Recchia and Louwerse, 2013). Furthermore, building high quality gazetteers is increasingly associated with merging gazetteer data from multiple sources, which in itself is a non-trivial task (Manguinhas *et al.*, 2008; Smart *et al.*, 2010). Finally, toponym usage in natural language is not restricted to official or administrative uses of placenames, and as such gazetteers should also contain both vernacular toponyms and the spatial footprints associated with them (Keßler *et al.*, 2009; Jones *et al.*, 2008).

4.3.1 Approaches to geoparsing

There are three basic approaches to identifying candidate georeferences: approaches which use (1) simple list lookup; (2) knowledge or rule-based methods; and (3) machine learning (Leidner and Lieberman, 2011).

The simplest, arguably the baseline approach that could be used to identify georeferences is to look up known entities from previously generated lists of place names, addresses, postcodes, phone numbers, etc. This approach is simple, fast and language-independent and has been used with success to identify locations in a range of texts. For example, Mikheev *et al.*, 1999 showed that performance for simple list lookup for locations could achieve precision of 90-94% and recall of 75-85% with 5,000 locations collected from the CIA World Fact Book and evaluated on MUC-7[4] data. They highlighted that the quality of the list also had a more significant impact than its size.

However, this simple approach suffers from a number of drawbacks: (i) the method is not able to identify new entities not found in the gazetteers (referred to as the *finiteness* of the list); (ii) a decision must be made about how to match with entries in the list (e.g., whether to allow partial matches or insist on complete matches); (iii) it is not always correct to assume that a word is being used in a geographic context, e.g., Chicago can represent the US city, the name of a pop group, or the internal project name for Windows 95 (McCurley, 2001); and (iv) georeferences often appear in different forms, e.g., "UK" and "United Kingdom", which will not match unless both forms are included in the list. Variants of names may also be used, such as historical or vernacular expressions (e.g., Zurich's Latin name of *Turicum* or the Scots vernacular name of *Auld Reekie* for Edinburgh).

A more effective approach is to make use of the surrounding context. For example, in the sentence "Keep up on your reading with audio books"[5] an approach using gazetteers may identify the words "on", "reading" and "books" incorrectly as locations in Vietnam, the UK and Louisiana, USA respectively. However, in this context the entries are clearly not being used in a geographical sense. To capture the surrounding context, rules are typically

[4]MUC stands for Message Understanding Conference - in the 1990s the MUC evaluations, organised by NIST in the USA, aided the development of metrics and algorithms to support evaluations of Information Extraction technologies, see: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/.

[5]This example is taken from the book "Natural language processing with Python" that describes NLTK, a publicly accessible toolkit for language processing in Python (Bird *et al.*, 2009).

defined and expressed within a grammar. Contextual evidence is often distinguished as either *internal* or *external*. Names often have some kind of internal or phrasal structure, which suggests they are names. This information can be stored or guessed and includes, for example, capitalisation (CapWord), prefixes or suffixes (e.g., company designators) and lists of names (the list lookup approach makes use only of internal evidence). For example, the internal evidence for a location might be captured as: "CapWord + Street, Boulevard, Avenue, Crescent, Road" and would identify expressions in text, such as "Portobello Street" and "Sunset Boulevard".

Additionally, there may be external (or contextual) evidence present within the text that makes it clear what type of entity a word or phrase is. For example, consider the phrase "President Washington chopped the tree". In this example "President" is clear external evidence that "Washington" is the name of a person and not a place. Rules can be built up to assist with distinguishing between classes of named entity, e.g., "North of CapWord" would normally help identify a place. In contrast a rule such as "Firstname + Location" where Firstname matches a list of commonly used people names and Location is a match in the gazetteer could be used to deduce that Location is not being used in a geographic sense in this context. The matching of text is performed through defining *regular expressions* that encapsulate the rules. Pattern matching tools such as *lex*, *flex* or *fgrep* are commonly used to apply regular expressions.

In older NER systems the rules were typically hand-crafted; however, most modern approaches make use of machine learning methods to induce rules automatically from previously classified (or annotated) training examples using *features* that capture contextual information. A training phase involving manually-annotated data is used to build a model based on the specified features. The trained model can then be used to classify new unseen examples. Machine learning methods include Hidden Markov Models (or HMMs), Decision Trees, and Maximum Entropy. By way of example, Curran and Clark, 2003 use a Maximum Entropy tagger to perform language independent NER because the tagger uses features which appear independent of the language. Features include information about Part-of-Speech (POS), a list of preceding and following named entities (NEs), orthographic information and whether words appear in lists of first and last names. On data for the CoNLL-2003 NE task, Curran and Clark achieve F_1 scores of 87.7% (English), 71% (German) and 83.2% (Dutch) for identifying locations.

Two issues of concern with the machine learning approach are: (i) the training data; and (ii) generalisation of the resulting classifiers. The first issue concerns how much data is sufficient for a machine learning approach to train a classifier, which can be used reliably on new unseen texts. Careful selection of training set examples is often necessary to create a balanced and representative training set. The second issue concerns whether the induced classifiers will generalise across new unseen texts or whether they only operate successfully

on examples found in the training set. This is often dependent on the features selected to train the classifier.

In the context of GIR, various authors have also discussed geoparsing web content (McCurley, 2001; Borges *et al.*, 2003; Morimoto *et al.*, 2003). For example, Borges *et al.*, 2003 use *wrapper induction* to extract addresses from web pages. Wrappers are IE systems which rely heavily on document formatting and markup language to induce patterns for IE, e.g., in HTML this includes access to tables and lists. Example pages with addresses annotated by users are used to create general purpose wrappers by deriving patterns based on the structure of the web page and HTML markup surrounding the address. A set of regular expressions are derived which can be used to extract further addresses from unseen examples. This form of rule generalisation (i.e., extracting rules, which rely not on the address itself but its structure and format) were successfully used to extract data from pages in the same web source, 65 pages from selected sites about hotels, restaurants, pubs, museums and other cultural attractions.

4.4 Geocoding: resolving georeferences

Having identified candidate georeferences, the subsequent task – referred to as *geocoding* or *toponym resolution* when carried out using only place names – is to associate them with a unique and, in the context of some knowledge base of geographic locations (e.g., a gazetteer or ontology), meaningful identifier, which ideally can be then associated with a metric georeference to a location (Buscaldi, 2011; Leidner, 2008; Leidner and Lieberman, 2011; Andogah *et al.*, 2012). For example, the toponym "Bath" in the south-west of the UK could be assigned geographic coordinates (51.3794N, −2.3656W).

A common issue with geocoding toponyms is that the toponym identified may refer to one of many locations with the same name (*referent ambiguity* or *geo/geo ambiguity*) - this is typically indicated by having multiple entries with the same name in a gazetteer. For example, the name "Chapeltown" refers to a location in South Yorkshire (UK), Lancashire (UK), Kent County (USA) and Panola County (USA). Analysing linguistic context can often allow the toponym to be correctly resolved. Even more structured information, such as addresses, is often subject to referent ambiguity outside some defined domain of reference (e.g., there is a Trafalgar Street in both London, England and Ontario, Canada). Another problem commonly encountered in the geocoding process is that different (or alternative) names may be used to refer to the same location, e.g. "London" and "Londres". In addition the alternative names may also be informal or vernacular names, e.g., "New York" can also be referred to as the "Big Apple". These examples of ambiguity are characteristic of natural language and geocoding on the whole can be likened to the linguistic task of Word Sense

Disambiguation or WSD – determining what dictionary sense a word in a text refers to (Ide and Véronis, 1998).

4.4.1 Approaches for geocoding

There are a number of approaches for geocoding (referent disambiguation or toponym resolution). The goal of a toponym resolution algorithm is to select a location from a set of candidate locations for each ambiguous toponym. Many methods rely on the use of external geographical resources, such as gazetteers[6] to map between the georeference and spatial coordinates and the use of heuristics and hand-built rules[7]. These heuristics will often come in the form of additional world knowledge about locations – such as population statistics, land surface area, preference for type of place (city over village) or part-of relationships – that can be used in helping to resolve ambiguity. Additional heuristics may also come in the form of known linguistic or geographic properties. For example, the *one referent per document* property states that the location of a toponym referred to throughout a text will likely remain the same (Leidner, 2008). This is similar to the Word Sense Disambiguation property in which words are assumed to have "one sense per discourse" (Gale *et al.*, 1992). Another commonly used property is that of Tobler's law (Tobler, 1970) that states that things that are closer together are more related to each other than things that are not. Approaches based on heuristics are often simple, efficient to execute and perform fairly well across multiple domains (Leidner, 2008).

Buscaldi, 2011 groups toponym disambiguation methods into three broad categories: (i) *map-based*: those involving computing spatial distances; (ii) *knowledge-based*: those exploiting external knowledge sources to find disambiguation clues, such as population statistics; and (iii) *data-driven* or *supervised*: those based on machine learning techniques.

Consider the example in Table 4.2 describing London (UK) and London (Canada) from Wikipedia. Assume the toponym London has already been identified as a location during the geoparsing stage; therefore, the next stage is to ground this in space using a geographical resource (gazetteer or ontology). For example, given access to the GeoNames[8] global geographical database then looking up London as an inhabited place results in 67 possible locations, including London in the UK, London in Canada and several London occurrences in the USA. A simple approach would be to assign ambiguous places a *default location* (or default sense) based on external world knowledge, for example the most commonly occurring place or largest population of a place (see, e.g., (Andogah *et al.*, 2012)). Many geographic resources, such as GeoNames, include additional information about locations;

[6] The gazetteers used are typically the same as those used in the geoparsing stage as it often only makes sense to extract georeferences that can be subsequently mapped to spatial coordinates.

[7] Leidner, 2008 provides an excellent review and analysis of geocoding methods and heuristics used in past work.

[8] http://www.geonames.org/

therefore, in the first example in Table 4.2 the largest populated London in GeoNames would be the correct one – London, UK. Approaches based on a default sense have been shown to work well (Li *et al.*, 2003); however, if the simple default approach was applied in the second example, whereby the same toponym is resolved to the same location regardless of context, the geocoding would be incorrect. Many approaches make use of default sense heuristics when all other methods of disambiguating the toponym fail.

Table 4.2: Example texts with toponyms to identify and resolve.

Example 1: London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames, London has been a major settlement for two millennia, its history going back to its founding by the Romans, who named it Londinium.
Example 2: London is a Canadian city located in Southwestern Ontario along the Quebec City – Windsor Corridor. The city has a population of 366,151 according to the 2011 Canadian census. London is at the confluence of the non-navigable Thames River, approximately halfway between Toronto, Ontario and Detroit, Michigan. The City of London is a separated municipality, politically separate from Middlesex County, though it remains the county seat.

Similar to geoparsing, the examples in Table 4.2 highlight the importance of context in determining the correct location. Many approaches to geocoding make use of contextual cues and one source of evidence is other place names mentioned within the same text segment (e.g., paragraph), or within the document as a whole. The occurrence of 'England' in the first example or 'Ontario' in the second are examples of *containment relationships*. Such information is commonly found in gazetteers, for example the Thesaurus of Geographic Names (TGN), as a hierarchical ordering of place names, e.g. `World > Europe > United Kingdom > England > Greater London > London` and `World > North and Central America > Canada > Ontario > London`. The appropriate location could be selected based on a text similarity score between the words in the hierarchy and the words preceding and following the ambiguous toponym or some measure of distance between items in the hierarchy (known as *ontological distance*[9] e.g., (Amitay *et al.*, 2004)).

The use of containment relationships is an example of a knowledge-based method (Buscaldi, 2011). However, a problem with relying solely on toponyms within a local or document context is that there may be insufficient occurrences of such names to enable successful disambiguation. An alternative approach would be to also take into account non-geographic names in the local context of a toponym (Overell and Rüger, 2008). Thus, Speriosu and Baldridge, 2013 train a supervised learning algorithm with geo-tagged Wikipedia articles and use all words in the surrounding document to perform text-based disambiguation. The idea is based on the assumption that specific non-geographic words will often occur in context with specific places. For example, in a text about Portland the occurrence of the

[9]For example, `World > Europe > UK > England > Slough` is distant from `World > North America > Canada > London` but relatively close to `World > Europe > UK > England > London`.

term lobster will indirectly suggest Portland in Maine rather than Oregon or Michigan. In that paper the authors show that some of the most effective results for toponym resolution can be obtained by learning the words associated with specific toponym instances based on those words in the immediate context of a disambiguated occurrence of the toponym. Thus when disambiguating an occurrence of an ambiguous toponym its context words can be matched against the previously generated language models of each instance of that toponym. This would be an example of what Buscaldi, 2011 refers to as a data-driven approach whereby contextual clues are captured as features and the problem is treated as a supervised learning or classification task. It is also an example of applying the language modelling methods that we discuss in Section 4.6.

Alternative approaches make use of spatial relationships rather than those based only on text (what Buscaldi, 2011 would class as map-based methods). These methods use the coordinates of places appearing in the context of an ambiguous toponym to perform disambiguation (in contrast to text-based methods). Here the assumption is that locations in a document are spatially autocorrelated, and thus the correct location will minimise distance to disambiguated referents (referred to as *spatial minimality*) (Smith and Crane, 2001). Leidner, 2008, for example, computes this distance for the positions of all candidate locations of the ambiguous name and selects, using a spatial minimality measure, the closest. In the second example in Table 4.2 then London will be (on average) spatially nearer to 'Southwestern Ontario' 'Quebec City - Windsor Corridor', 'Toronto', 'Ontario', 'Detroit' and 'Michigan' for the correct location in Canada compared to the locations mentioned in the context of London, UK. Particularly effective are instances of unambiguous toponyms which provide a stronger form of evidence against which to ground ambiguous toponyms (Li *et al.*, 2003; Rauch *et al.*, 2003). Results from experiments by Speriosu and Baldridge, 2013 show that their data-driven approach based on using text-based methods outperform baselines for random toponym selection, resolving based on the largest population of a place and standard minimality-based resolvers (i.e., map-based approaches).

Further issues to consider based on the examples in Table 4.2 would be how to ground features that cannot easily be assigned points (e.g., the *River Thames*) or dealing with historical references to place (e.g., *Londinium*). Smith and Mann, 2003, for example, show that geocoding is much less successful for historical documents compared to current news articles. Further geocoding challenges are encountered when geocoding at finer levels of granularity. For example, Pasley *et al.*, 2007 show that street-level data exhibit higher levels of ambiguity compared to those at region and city level. In practice it may also be common to make use of multiple geographical resources and therefore an additional step of aggregating the resources (based for example on similarity of coordinates and of the names and feature types) or resource selection may also be necessary (Smart *et al.*, 2010). Clough, 2005 describes an approach in which higher quality resources provided by national mapping agencies are preferred when disambiguating placenames for specific countries compared to

using resources with global coverage but variable quality (e.g., GeoNames) (Graham and De Sabbata, 2015; Ahlers, 2013).

4.5 Computing document scope

Andogah *et al.*, 2012 define the geographical scope of a document as "the geographical regions or areas which the document is about" (p. 1). The process of *geographical scope resolution* is the automated assignment of geographical scopes to a document. This follows the intuition that many documents, particularly web pages, are likely to be of relevance to people within a specific (local) region in contrast to more general (global) ones. For example, the BBC news pages about the UK may be of higher relevance to people from the UK rather than worldwide; the relevance of Sheffield City Council web pages may be greater to people from Sheffield, etc. One of the earliest papers to discuss geographic scope was Ding *et al.*, 2000 who computed geographical scopes of web resources, such as web pages, to establish which audience they are targeted for – residents of a city, country or the world. Textual content of web pages together with the geographical distribution of hyperlinks are used to estimate geographical scope.

Wang *et al.*, 2005 further categorise the locations of web resources into three types: (i) *provider location*; (ii) *content location*; and (iii) *serving location*. Provider location describes the physical location of the provider (e.g., organisation or person) owning the web resource, e.g., identifying service providers from Yellow Pages within a specific region (Himmelstein, 2005). Content location describes the location the content of the web resources is about. Finally, serving location describes the geographic scope that the web resource reaches. For example, information about paying council taxes provided by Sheffield City Council may be only relevant to residents of Sheffield; information from the UK Government website about visas would have a more global reach. In practice, however, many applications appear simply to either define document scope based simply on a dominant location or a weighted mean of all the locations in a document.

Computing scope is often based on ranking and selecting locations identified within the document, such as the most frequently occurring georeference. Many of the exemplar systems involve a process of identifying appropriate document scopes, including Web-a-Where, NewsStand and STEWARD (see Section 4.8).

4.6 Modelling locations implicitly through language modelling

The development of a wide range of resources associated with explicit coordinates, the increasing maturity of machine learning approaches, and the need for generalisable methods

applicable to very large volumes of data in real time has led to a new family of methods which rather than exploring toponyms explicitly contained within text, seek to learn ways in which location is described more generally in text. The key insight is that given numerous documents associated with coordinates, such as geotagged Flickr captions, Wikipedia or Tweets, it is possible to identify sets of words that are associated with particular regions of space. The set of frequencies of words for a given region is referred to as a language model. An early example of applying the idea is in (Ahern *et al.*, 2007), who used k-means clustering of coordinates in combination with a modification of tf – idf to select significant keywords from Flickr tags, that were then allocated to the cells of a latitude/ longitude grid, at different levels of granularity (cellsize). This formed the basis of a browsing system for geographic exploration of geographically specific concepts (as identified by the language models) and the associated photos. Other early applications were concerned with geo-referencing Flickr images that have no coordinates (though they use geocoded images to evaluate the methods). The main idea is to match the tags of a photo to the most similar language model and hence geographical location. This is sometimes done using Naive Bayes machine learning methods. The language models can be learned from geocoded Flickr tags or from other geocoded text. Crandall *et al.*, 2009 adopted a spatial subdivision based on mean shift clustering and used both Bayesian and Support Vector Machine (SVM) machine learning methods to associate an image with a location using evidence from both textual tags and from visual (SIFT (Lowe, 2004)) features in the images themselves.

O'Hare and Murdock, 2013 addressed the same problem of geocoding Flickr photos using Bayesian methods with only textual evidence for language models associated with regular (latitude / longitude aligned) grid cells at multiple scales. The decision on the cell to which a photo was allocated also used the language models from grid cells neighbouring the target cell. Kinsella *et al.*, 2011 applied language modelling methods to locate Tweets and their users, and experimented with matching the tweets to locations using Kullback Leibler divergence (distance) as well as Bayesian probability. They found that at the finest levels of granularity the language models were clearly superior to gazetteer based methods of geo-coding (using Yahoo! Placemaker), apparently due the lower likelihood of users including place names in tweets referring to very localized situations.

A notable example of applying the language model approach to attach coordinates to more substantial documents, i.e. Wikipedia articles, is that of Wing and Baldridge, 2011. They again used a regular grid to partition the language models and estimated the probability of a document belonging to a 1 degree geographic cell based on the Kullback Leibler divergence between the document language model and the cell language models. They also investigated Bayesian methods. The cell language models were built from the text of all geocoded Wikipedia articles that fell within the cell. Roller *et al.*, 2012 developed the method further, by using a clustering method that provides different sized cells, and by modifying the final matching step to select the centroid location of the geocoded Wikipedia

articles from the training set in the selected cell, rather than just the centre of the cell. Laere *et al.*, 2014 improved significantly on these results for geo-coding Wikipedia articles by using Flickr, Twitter and Wikipedia to create richer language models, with k-medoid clustering, and by selecting the (geocoded) resource in the training data that was most similar in its textual content to the document being geocoded. The higher quality results were obtained when geocoding Wikipedia articles that related to "spot" i.e. fine grained, locations such as buildings, as opposed to spatially very extensive objects such as rivers.

Typical approaches to evaluating these approaches use a test dataset of resources with known locations and measure mean and median distance of geocoded documents to their actual positions. In a recent review, Melo and Martins, 2017 reported on results from a wide range of approaches, where median values ranged from 2.2-640km and mean values varied between 83-2854km for English Twitter and Wikipedia datasets. Clearly these results are quite coarse-grained and while they show good potential, with some very good results, it appears at the time of writing that they might be best suited to geocoding resources that have no or very few locationally specific toponyms (such as some micro-blogs) or to summarising the geographic distribution of an entire collection of documents.

The language modelling approaches have however been shown to result in considerable performance benefits for purposes of toponym resolution, as demonstrated in the work of Speriosu and Baldridge as described earlier. Building on a similar approach, DeLozier *et al.*, 2015 used language modelling methods to georeference toponyms that are not present in gazetteers. Thus the context words (e.g., the 10 words before and after), of a word recognised using NER methods as a likely toponym, are each associated with a probability of occurrence at locations on a grid, and the combination of these density fields for all context terms is used to identify the grid location that is most likely to be associated with the toponym. If the toponym is also present in a gazetteer then this approach allows the candidate location to be selected as the one that lies in, or is closest to, the respective grid cell. This latter approach, which is called TopoCluster, was shown by DeLozier *et al.*, 2015 to be superior for toponym resolution of gazetteer names to the methods presented in Speriosu and Baldridge. A recent paper by Gritta *et al.*, 2017 compared several state of the art approaches for toponym resolution (and toponym recognition) including TopoCluster, The Edinburgh Geoparser (Grover *et al.*, 2010), GeoTxt (Karimzadeh *et al.*, 2013), Yahoo!PlaceSpotter[10] and CLAVIN[11]. For toponym resolution with their Wikipedia-based corpus (created using both Wikipedia and GeoNames) they found that both Yahoo!PlaceSpotter and The Edinburgh Geoparser performed better (according to all their evaluation metrics) than TopoCluster. On the toponym recognition task TopoCluster performed very well but not as well as the front runner for that task, which was The Edinburgh Geoparser.

[10] https://developer.yahoo.com/boss/geo/docs/key-concepts.html
[11] https://clavin.bericotechnologies.com/about-clavin/

4.7 Evaluating geoparsing and geocoding

Given the importance of georeferencing, a key task is evaluating its quality. Despite this, surprisingly few resources exist for evaluating geoparsing and geocoding methods, and hence providing gold standard data against which different approaches can be developed, evaluated and compared. One reason is the large amount of work involved in the manual analysis of a collection of documents selected as a representative sample of texts. All geographic references must be identified and demarcated using a suitable annotation scheme and format (e.g., XML). Additionally, to evaluate geocoding the correct location must be identified for each toponym, either specifying spatial coordinates or mapping to the correct entry in a gazetteer or ontology. The annotations can then be measured and compared against the output of an automated system.

To evaluate the success of geoparsing we must compare two sets of annotations: a manually-generated set (*key-set*) and a system-annotated version (*response-set*). Various measures of annotation overlap can be computed from a contingency table containing information about correct annotations (C) as well as false positives (FP) and missing (M) annotations. These include standard IR measures such as precision (P), recall (R) and the F_1-measure (F_1). From the annotations in the response-set, precision measures the proportion that correctly match the manually assigned annotations. From the annotations defined manually, recall measures the proportion of these which are also correctly identified by the geoparser. The F_1-measure score is a single-valued summary of both precision and recall and enables much simpler comparison of different geo-parsing methods. Precision, recall and F_1 are commonly used measures in the evaluation of Information Retrieval systems and text categorisation algorithms. Many resources exist to evaluate NER systems more generally and beyond location. For example, resources from the Message Understanding Conference (MUC), the ACE evaluation and the CoNLL 2003 Shared Task (see (Leidner and Lieberman, 2011) for more details). It is important to note that the use of a single measure (e.g., precision) can be significantly biased by a corpus - for instance simply using default sense and predicting that all instances of London are associated with London, England is likely to achieve high precision, but fail to identify any other instance of London correctly.

As well as using measures based on contingency tables, it is also possible to explore the proportion of toponyms whose location is the same as some reference data, and to quantify error through metrics, such as Root Mean Square Error (RMSE), where the location of a georeferenced toponym is considered as a quality measure (Leidner, 2008; Wing and Baldridge, 2011; Palacio *et al.*, 2015).

4.8 Georeferencing in the exemplar systems

In the following we briefly describe the approach taken by the exemplar systems introduced in Chapter 1.

The Web-a-Where system implements a gazetteer-based approach to assign a geographical scope to the target geography of a web page, in other words the geography discussed within the content of the page, and identifies and disambiguates cities, states, countries and continents in the text and assigns locations. Each location takes the form of an entry from the gazetteer, which is structured in a hierarchy of other geographical references that encompass it (e.g., `Country>State>City` for cities in the US) and used also for disambiguation. This containment hierarchy is created using existing resources, including GNIS and world-gazetteer.com, but requires the population of countries and cities to be over 5,000, resulting in an integrated gazetteer of approximately 40,000 locations. This also includes lists of entries commonly used in a non-geographical sense to aid disambiguation (in step ii below). Web pages are processed in three stages: (i) identify possible names in the text based on scanning the gazetteer; (ii) assign a location (using confidence levels) to each possible name; (iii) determine the focus of the web page as a whole (constrained to city, state, country or continent). Web-a-Where makes use of heuristics to disambiguate place names and identify the focus of a page and therefore illustrates a rule-based approach. Heuristics include one sense per discourse, default sense (based on population) and the co-occurrence of placenames within the same context. The containment hierarchy is also used to resolve ambiguous placenames based on proximity within the document and also their spatial proximity as deduced from the hierarchy.

The approach used in the SPIRIT system is also constrained to matching entries from a gazetteer and uses various heuristics (e.g., default sense) to disambiguate georeferences from web pages (Clough, 2005). A rule-based approach that makes use of local context is implemented using the General Architecture for Text Engineering (GATE) tool. Similar to Web-a-Where multiple geographical resources with worldwide and country coverage are combined into a single knowledge-base. Quality resources from national mapping agencies were used to provide greater coverage for the countries addressed in the project: UK, Switzerland, Germany and France. Similar to Web-a-Where the gazetteer contained hierarchical information for each entry that was used in the geocoding stage. For geoparsing, matches with entries in the gazetteer are first identified with rules in the form of a grammar used to filter out those unlikely to be used in a geographical sense, e.g. `<title><placename>` would filter out 'Mr Bath'. Names commonly used in a non-geographical sense (e.g., 'of') were removed from the gazetteers. Although this would harm cases where common words are used in a geographical sense, results overall demonstrated significant improvements in performing this step. The SPIRIT geotagger could also identify and ground toponyms and postcodes from addresses. In the case of ambiguous placenames various approaches

were used including a default sense, resource preference and matching terms from the local context and entries in the hierarchy of matching entries from the gazetteer. Similarly, Wang and Stewart, 2015 adopted GATE's standard tools by enhancing the standard gazetteer used by GATE for their work on natural hazards.

The FrankenPlace system makes use of multiple gazetteers, such as GNS and the Getty Thesaurus of Geographic Names (TGN), and combines these into an integrated gazetteer (Adams *et al.*, 2015). Existing tools are used to perform geotagging as this was not the focus of the work. These included the CLAVIN[12] (Cartographic Location And Vicinity INdexer) tool, an open source software package for document geotagging and geoparsing that employs context-based geographic entity resolution. The GeoNames web service is also used to match placenames to GeoNames ids.

In the STEWARD system an approach utilising Part-of-Speech (POS) tagging and NER is used to perform the geoparsing step (Lieberman *et al.*, 2007). Unlike the approaches used in Web-a-Where and SPIRIT, gazetteer lookup is performed *after* identifying potential placenames. POS tagging is used to identify proper nouns first and then NER applied to them along with contextual information to aid the NER process. Only proper nouns identified as locations by the NER system are further considered. The result of this step is a feature vector of possible locations. Geocoding is performed by comparing the possible locations with entries in the GNIS gazetteer (2.06 million locations in the US). For an ambiguous placename all gazetteer entries are extracted and disambiguation makes use of the property that authors will commonly mention nearby placenames or more identifiable geographic locations in the same textual context as the ambiguous one to aid readers less familiar with a region (as is often the case in news reports). Pairs of co-occurring names are then compared and the most likely gazetteer entry selected based on the 'strength' between them. This is computed based on the frequency of occurrence, document distance (the difference in offsets of the locations from the start of the document), geodesic distance and the populations of the pair of locations. The algorithm computes all possible pairs and ranks the matching gazetteer entries in order of strength with the strongest selected as the most likely location. Finally the scope of each document is computed based on ranking the geographic locations mentioned in the document. The frequency of occurrence could be used to sort the locations, but in the case of STEWARD an algorithm called *Context-Aware Relevancy Determination* is used that treats mentions of toponyms which both occur throughout a document, and are contextually related (that is frequently co-occur within a document) to spatially proximate toponyms, as being more important in determining scope.

The NewsStand system built on work in STEWARD, and again used a mixed set of methods based around gazetteers (Lieberman *et al.*, 2010), where a distinction is made between a *global lexicon* of locations known to all audiences, and an audience-specific *local*

[12] https://github.com/Berico-Technologies/CLAVIN

lexicon. For example, Paris in France is typically known to everyone; whereas Paris in Texas would be more familiar to people familiar within that region. This aspect of the geotagging approach distinguishes it from most other systems, e.g. Web-a-Where and SPIRIT that make use of global lexicons. The GeoNames gazetteer comprising over 7 million entries is used to validate potential toponyms found using geographic cue words (e.g., 'city of X') and Part-of-Speech tagging. A number of heuristics, such as geographic proximity, the use of local lexicons before global and one sense per discourse are used for toponym resolution. Lieberman and Samet, 2012 make use of adaptive context features to improve geotagging in the case of streaming news, e.g. from Twitter. The adaptive method is compared with the output of existing web-based services for geo-tagging: the OpenCalais[13] and Yahoo! Placemaker[14] systems.

Both projects associated with fine-granularity toponyms describing mountain regions (Gaio *et al.*, 2008; Derungs and Purves, 2014) used context specific rule-based approaches and fine granularity gazetteers to georeference content, with the PIV project explicitly dealing with representing more complex geographic regions found in text (e.g., between London and Edinburgh) while Derungs and Purves incorporated a Digital Elevation Model[15] in the context information they used in georeferencing. At a very different scale, Brown *et al.*, 2012 linked context to gazetteer data using machine-learning approaches, and generated not document scopes, but regional topic models.

4.9 Summary

This chapter has discussed approaches to identify potential references to location in text (georeferenceas) and their subsequent resolution to a location on the earth (geocoding). Both tasks require the use of contextual information in order to disambiguate and ground geographical references and enable their use in GIR systems. We also discussed related concepts such as computing document scope and how the exemplar systems have in practice analysed text documents for references to place. We note that despite the promise of machine learning-based methods, most of the differences in our exemplar systems appear to result from customisation of an approach to a particular domain, for example in terms of the type of language expected, the nature and richness of gazetteers and the granularity required from the georeferencing process. Thus, it appears, at least for the moment that, as well as methodological expertise, detailed domain knowledge remains an indispensable part of the toolset of any researcher working on the georeferencing process (Leveling, 2015), and its importance should not be underestimated.

[13] http://www.opencalais.com/

[14] http://developer.yahoo.com/boss/geo/

[15] A Digital Elevation Model is a regular grid of heights (a field in the language of Chapter 2 from which a range of properties can be derived including local relief, gradient and aspect

5 Indexing

5.1 Introduction

The index is the component of a GIR system that contains references to all documents known to the system, either as a result of crawling the web or being input to the index from other data sources (Section 5.2). In this chapter we discuss common approaches to indexing, including the use of inverted lists (Section 5.3), spatial indexing (Section 5.4) and combinations of textual and spatial indexes (Section 5.5). We finish the chapter by discussing indexing methods in our exemplar systems (Section 5.6).

5.2 The need for indexing

The purpose of the index is to provide fast access to documents that match the spatial and thematic components of a query. Indexing structures that provide access to documents that contain particular terms are called inverted lists, also referred to as an inverted file or inverted index (Manning *et al.*, 2008). GIR systems often use inverted index techniques, but they are combined with spatial indexing methods to determine which documents relate to specific regions of space.

Because speed of access is a major concern, it is important that the index provides access to all the main data items that are required to rank the retrieved documents. Thus there could be millions of documents that match the textual and spatial query specification, but the user usually only wishes to be presented with those few documents most likely to be of interest, as determined by the relevance ranking procedure (see Chapter 6). Data items required for ranking can be stored in the index, and they could include (among many others) the number of documents that include a particular term, the number of times a term occurs in a specific document and where in the document the text term actually occurs (see Section 5.3).

In practice, indexes for web search refer to billions of documents and so various strategies are employed to speed up the retrieval process. For example, one or more index structures are used to perform an initial retrieval of the set of documents that contain at least one of the query terms. This is an extremely fast process as it is a simple boolean match, and it greatly reduces the number of documents that are ranked. The smaller resulting set of documents can be refined by other indexing structures that provide access to data items used to perform relevance ranking, with respect for example to the distance from the user's location and the personal profile of the user. There may be several layers of indexes and caches, to improve both the speed of the retrieval as well as the quality of the results. Personalisation and spatial information typically are incorporated in the ranking over a small subset of the documents higher in the architecture because they are computationally

Term	List of Document Ids.
Term 1	D1, D3, D6, D8
Term 2	D2, D3, D7
Term 3	D1, D2, D7, D9
Term n	.....

Figure 5.1: A simple inverted file in which each record contains a term and a postings list with references to the documents in which the term occurs. In practice there are many other items of data that may be stored in the index.

more complex, and the results cannot usually be cached for other users. Often the index will include structured metadata identifying the geographical scope of a document, and other salient information useful for ranking, which has been extracted at indexing time, or linked from other data stores. Because of the vast quantities of data and the need for fast access in web search engines, an index may be duplicated or split into parts to support distributed processing (Manning *et al.*, 2008). This allows processes to be farmed out to multiple computers, which both increases speed and provides resilience to machine failures.

We do not address the methods that are used in commercial search engines in detail because they are proprietary and not replicable, subject to constraints such as license agreements, and often do not represent the state of the art in the literature. In the remainder of this chapter we briefly review the inverted text file indexing methods and summarise methods for spatial indexing, before providing a summary of the various approaches that have been proposed for their integration.

5.3 Indexing with inverted lists

In the standard method of indexing documents, with an inverted list, the index consists of key:value pairs, where the *keys* are the set of all terms (mostly words), i.e. the dictionary or lexicon, that occur in the documents and for each term there is a list of the documents (the *value*) that contain the term, referred to as the postings list (Figure 5.1). The dictionary can contain stemmed versions of the terms, but it may also keep track of the original version of the word to allow precise matching between query terms and relevant documents. In order to allow the index to support access to quoted phrases, the posting lists need to include data on where in the document the respective term actually occurs. To support relevance ranking procedures, the index will also store other data such as the number of documents that contain a term, and for each document the frequency of occurrence of the term in the document.

Given a query in the form of a set of search terms, the inverted list is used to find, for each query term, the documents that contain the term. This is a fairly simple process in that each query term is matched to the corresponding key in the index and the posting list for that key is returned. To find the documents that contain all search terms, the intersection

of the posting lists corresponding to each search term is computed. To find documents that contain a specific search phrase, further processing is required to check whether the terms occur in sequence in the documents that result from the intersection operation.

5.3.1 Limitations of inverted indexes in GIR

Inverted indexes play an important role in GIR in that they enable access to documents that contain the concept terms of a user's query. They can also be used to find documents that contain the toponyms in geo-queries, as by default the toponyms will be treated exactly the same as all other terms in the indexing process. For some geo-queries this may work quite well and it was the approach adopted in the earliest examples of geographical search engines such as (Ding *et al.*, 2000) in which the emphasis was on determining the geographical scope of web documents rather than the problems of indexing.

There are several situations in which simply retrieving documents that contain the toponyms in a query will not suffice. One such situation is when a relevant document uses an alternative toponym to the one employed in the query. This can arise as many places have more than one name, including different language versions. It could also be that a document refers to places contained in the place denoted by the query toponym, without actually mentioning the parent toponym. Another situation that causes problems for this approach is when the query includes a spatial qualifier, such as *within X miles of*, *outside of* or *neighbouring*, in which cases the user is interested in a region that is related to the named toponym but is not equivalent to it.

An apparent solution to the problem of the document containing an alternative toponym to that of the query is to use a form of query expansion in which the query toponym is supplemented by alternative versions of the name. For anything other than quite small regions this can require a very large number of query terms, which could adversely affect the performance of the query processor. The latter criticism also applies to the possibility of generating toponyms that correspond to an interpretation of the region implied by a spatial qualifier.

5.3.2 Benefits of combining spatial indexing with inverted lists

The limitations of purely toponym based query led to the development of alternative approaches to indexing that combine inverted list indexing with spatial indexing techniques developed in the context of GIS. Here the geographic context of the document is represented not by a set of one or more toponyms but by regions of space represented by geometry such as a rectangle or polygon, or in some cases by just a single point. Thus, provided a document can be represented by a geometric *document footprint* that could be based for example on the geographic references present in the text of the document, and the query is also converted to a geometric *query footprint*, which corresponds to the interpretation

of queries such as *within 10km of place X* or *in place X*, then retrieval of geographically relevant documents is no longer dependent on an exact match between toponyms in the query and those in the document. Instead it is performed using spatial data processing that matches the query footprint to the document footprints. The geometry of the query footprint representing these spatial relationships can be computed using GIS methods such as those presented in earlier chapters.

5.4 Spatial indexing

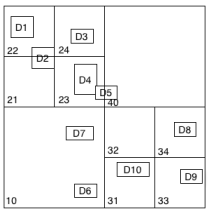
The purpose of a spatial index is to assist in retrieving those stored geo-referenced objects that lie in the query footprint. In Geographical Information Systems (GIS) the objects to be retrieved might be topographical features, for example buildings, roads or rivers, that are referenced to geographic space with geometry such as points, polygons and lines. In a GIR system the objects to be retrieved could be the documents or some other media that again are referenced to particular locations via the geometry of their document footprints.

There are two common approaches to building a spatial index, referred to as space-directed and object-directed. In both cases the index can be regarded as set of *key:value* pairs, where the *key* represents a spatial cell that either contains or intersects the geometry of the stored data objects referred to with the *values*. Access to the *key:value* pairs is provided either by a standard database indexing data structure such as a B^+-tree or by an indexing data structure that is specifically adapted to multi-dimensional (and hence spatial) data. A B^+-tree consists of a hierarchy (tree) of nodes in which the non-leaf nodes contain a set of keys and pointers to child nodes, while the leaf nodes store references to the stored data (which could be geometry objects), i.e. the *values* of the *key:value* pairs. The tree is balanced in the sense that all branches have the same number of levels.

5.4.1 Space-directed spatial indexing

In space-directed indexing methods, the key cells are typically rectangular in form and serve to tessellate the entire region to which the database refers. In the simplest form of space-directed indexing the cells are on a regular grid and fixed in size. Each cell is identified by a location code which is a number representing a point or a region of space. The codes are some function of the coordinates of the origin (corner) of the cell and in some cases also of the extent of the cell. These location codes can then serve as the keys in a B^+-tree [1].

[1] A desirable property of location codes is that codes (numbers) that are similar represent locations that are close together, so that a sequence of location codes should represent a contiguous region of space. In practice actual coding schemes such as Morton codes (also called Z-curves) and Hilbert curves vary in the extent to which they meet this ideal and will always have some successive numbers that correspond to discontinuities in space.



Cell ID	Document IDs
10	D6, D7
21	D2
22	D1, D2
23	D4, D5
24	D3
31	D10
32	
33	D9
34	D8
40	D5

Figure 5.2: A linear quadtree spatial index. Each cell is associated with a list of the documents whose footprints intersect the cell.

While the regular grid method is often used in practice, it has the disadvantage that there is no constraint on the number of data objects that might need to be linked to the cell, which can cause performance overheads. To counteract that, the cells may differ in size, as in quadtree spatial indexing schemes, in which cells are recursively subdivided until they contain or intersect no more than a specified threshold number of data objects (Figure 5.2). The individual cells in a quadtree index are usually identified by location codes, again based on the coordinates of the corner and possibly the size of the cell. A simple method to compute the location codes is by interleaving the bits of a binary representation of the coordinates, resulting in Morton codes. When access to the quadtree location codes is provided by a B^+-tree, the scheme is described as a linear quadtree (as opposed to a point-based quadtree that could be maintained in a tree data structure with four children per node). In Figure 5.2 the referenced data objects labelled D1, D2 etc are documents, the footprints of which intersect their respective cell.

5.4.2 Object-directed spatial indexing

For object-directed indexes the cells are also typically rectangular but their dimensions are adapted to fit around the geometry of individual stored objects, with the result that in such schemes the cells may overlap each other. In the R-Tree, the most common form of object-directed index, the bounding rectangles are organised into a hierarchy, so the lowest level rectangles are grouped into larger containing rectangles, which themselves may be contained by higher level rectangles (Figure 5.3). The keys in an R-tree are the dimensions of the bounding rectangles and the indexing structure is very similar in design to a B^+-tree, with the leaf nodes again containing references to the stored geometry.

5.4.3 Filtering and refinement stages of spatial indexing

When querying with a spatial index the region represented by the query footprint needs to be matched to the cells in the index - a process that can be treated as having two stages. Initially the cells that cover the query footprint are identified, so the stored data

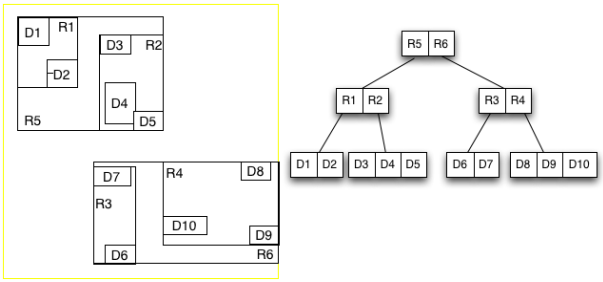


Figure 5.3: An R-tree spatial index. Each leaf node contains entries for the documents whose footprints are contained in the corresponding rectangle.

objects linked to those cells can then be retrieved. This is called the filter stage. Because the accessed index cells must completely cover the query footprint, and hence possibly extend beyond it, they could reference some stored objects that are outside the boundary of the query footprint. The second stage, called the refinement stage, performs a more precise test to determine which of the initially retrieved objects actually lie inside or overlap the boundary of the query footprint.

5.5 Spatio-textual indexing

There are various ways in which spatial and textual indexing can be combined, ranging from maintaining two completely separate index structures to various levels of integration.

Separate spatial and textual indexes

A simple approach to combining spatial and textual indexing methods, referred to as spatio-textual or spatial keyword, indexing, is to maintain separate spatial and textual indexes and merge the results for a particular query. Here the spatial index is used to retrieve documents that relate to the spatial constraints of the query, while the inverted index retrieves the documents that contain the concept terms of the query (which might or might not include place names). The results can then be intersected to find those that meet both the spatial and textual requirements either completely (as in AND semantics) or partially (as in OR semantics in which not all query criteria might be met). Relevance ranking can be performed by combining text-based and spatial relevance scores. This simplistic approach has benefits with regard to index storage space but can have overheads with respect to processing time when compared to schemes that provide closer integration (see (Vaid *et al.*, 2005)), in which as in several other papers, a number of different approaches are compared).

In Vaid *et al.*, 2005 the spatial index stores, for each spatial cell, the document IDs (docIDs) of documents whose footprints intersect the cell. Chen *et al.*, 2006 introduced some

alternative approaches to the use of a regular grid with separate inverted list, in which index storage costs were reduced by not referencing the docIDs explicitly from the spatial index cells. In their schemes, the main spatial index is a regular grid, of size 1024x1024 or 256x256. Each document footprint is represented by one or more "toeprints" where each toeprint corresponds to the Minimum Bounding Rectangle (MBR) of each region, such as a city, that the document refers to. These MBRs are themselves described by a range of location codes based on space filling curves such as Morton codes (Z-curves) and Hilbert curves (see section 5.4.1), the range being the longest interval of location codes that intersect the MBR. The toeprint ID is the number (identifier) of the toeprint when ordered by its location codes. Chen *et al.*, 2006 describe several methods for processing the data. In the more time efficient Tile Index, the cells of the regular grid maintain references to the toeprint IDs (rather than the docIDs). On retrieving the toeprint IDs for spatial index cells that intersect the query footprint, they are translated, via a separate data table, to the document IDs of those documents that refer to the corresponding region of space. These document IDs are then used as a filter to access the inverted list of document IDs that contain the query text terms. The toeprints associated with the resulting document IDs are then used to refine the set of document IDs that meet the spatial constraints of the query. This latter step is needed as some of the toeprints returned by the relevant cells of the regular grid index could be outside the query footprint.

Ontology-based spatial indexing

In another example of separate spatial and textual indexing, Brisaboa *et al.*, 2010 create a form of hierarchical spatial index in which the nodes reference named places rather than arbitrary spatial cells governed by the component geometry data items. The nodes are organised in a geographical hierarchy, though nodes are only present for places (or their parents) represented in the document collection. Each node is associated with a list of the documents that relate to the named place and with the minimum bounding rectangle of the place. The text content of the documents is accessed with a separate inverted file. There is also a hash table that maps place names to the storage address of the corresponding node in the geographic hierarchy. The benefit of this approach is that it provides a simple method to retrieve all documents that relate to a named place and all of its contained places, while also supporting spatial and textual queries and their combinations. The hierarchy of places is however not a balanced tree, as in an R-tree, and therefore lacks its beneficial properties with regard to query access times.

Spatial primary methods

An approach to closer integration of spatial and textual indexing is to use either space or text indexing as a primary filter for the query and to include, as part of the integrated

indexing structure, a secondary index that filters by the other, either textual or spatial, query features. This leads in the first instance to a spatial-primary index in which each spatial node of a spatial index (corresponding to a cell or region of space) is associated with an inverted list of the documents that relate to that spatial node (in that their document footprints intersect the space represented by the node). So given a query consisting of text terms T and a spatial footprint S, the inverted list associated with each cell of the spatial index that intersects S is searched to find those documents that include the query terms $t \in T$. Spatial primary methods have been implemented with a regular grid spatial index (Vaid *et al.*, 2005) and with an R-tree as in Zhou *et al.*, 2005, who demonstrated the expected relative benefits of using R-trees relative to a regular grid. The approach is equally applicable to a quadree. Figure 5.4 illustrates the idea with respect to a quadtree and an R-tree. Note that for each document that is referenced in the postings lists of the inverted index structures, its footprint(s) must be accessible at query time in order to check that they do actually intersect the query footprint, i.e. to perform the refinement step referred to earlier in the context of spatial indexing. Thus the fact that a query footprint intersects an R-tree rectangle or a quadtree cell does not guarantee that it intersects all the footprints of the referenced documents.

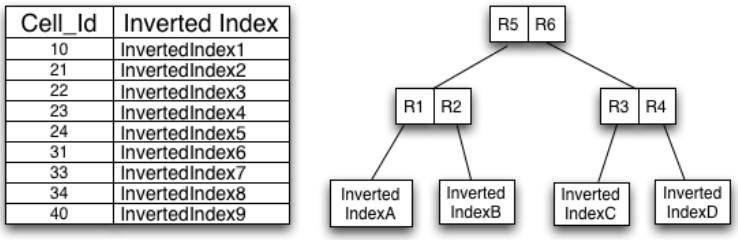


Figure 5.4: Spatial first spatio-textual indexing with a quadtree (left) and an R-tree (right). Note that the inverted indexes referenced by the nodes of these structures refer just to those documents whose footprints intersect that node. Data must also be stored for each document footprint so that it can be tested against the query footprint.

There have been several variations on the basic idea of spatial primary indexing. Thus in the KR*-tree (Hariharan *et al.*, 2007a) all R-tree nodes, including non-leaf, store references to both spatial rectangles and the keywords that appear in the respective rectangles of that node and its children. Keywords refer here to the concept terms that characterise the geo-datasets that the authors indexed. They are analogous to the terms that occur in text documents. The keywords point to the child nodes referring to data objects that contain them. Both space and text terms are considered simultaneously as the index structure is traversed, hence reducing the number of irrelevant nodes that will be examined in the tree. They also investigated storage of an inverted list structure at a selected higher level of the tree as an alternative to keyword storage at each node. The authors claimed fewer disk

accesses in comparison with the R-tree methods of Zhou *et al.*, 2005, but their experiment was conducted with geo-datasets which can be expected to have many fewer text terms than would be found in normal web documents.

Retrieving top-k most relevant documents

The idea of storing text terms at nodes of an R-tree is also found for example in the IR^2-tree (De Felipe *et al.*, 2008) and the IR-tree (Cong *et al.*, 2009). Both of the latter schemes support retrieval of the top-k most relevant documents using a weighted combination of both space and text terms (as in most other spatio-textual relevance ranking schemes), and represent spatially both the query and individual documents by a point. Both are notable for building upon the incremental nearest neighbor algorithm for R-trees introduced in Hjaltason and Samet, 1999. Note that the previously mentioned schemes retrieve all documents that contain the query terms and are within the query footprint, prior to relevance ranking and are hence less efficient. In the IR^2-tree all nodes maintain a signature file that records distinct text terms associated with documents referenced by nodes of the respective subtree. This auxiliary data structure is smaller than and improves upon the analogous node-specific data structure of the KR*-tree (Hariharan *et al.*, 2007a) described earlier.

In the IR-tree (Cong *et al.*, 2009) each leaf node of the R-tree points to an inverted list of its corresponding documents along with the weight of each term within its respective document. Each non-leaf node also includes a reference to an inverted list (called a pseudo-document), for the text terms in that subtree, and each term is associated with its maximum weight in documents of the respective child node. In a variation of this IR-tree, called the DIR-tree, the subtrees of the R-tree are optimised not just for clustering in space, as in a conventional R-tree, but insertion in the subtree is also a function of textual similarity of the referenced documents. This has benefits in query times for top-k queries, especially when the number of query terms is increased. Variations on the IR-tree of Cong *et al.*, 2009 to improve clustering of data are described in Wu *et al.*, 2012 who also introduced a top-k ranking procedure in which the query is represented by a region rather than a point. Their top performing access method was the CD-CDIR-tree.

Li *et al.*, 2011 describe an indexing method also called an IR-tree that is very similar to the IR-tree of Cong *et al.*, 2009 except that document summaries of non-leaf nodes (equivalent to pseudo-documents of Cong *et al.*, 2009) are maintained in a single inverted file. In Wu *et al.*, 2012 variants of the IR-tree of Cong *et al.*, 2009, including the CD-CDIR-tree, are shown to be superior relative to Li *et al.*, 2011 with respect to query time but not regarding index size.

Text primary methods

In the alternative view, of a text-primary index, the documents in the inverted list for each text term are organized by a spatial index dedicated to the respective term (Figure 5.5). In that case, for each text term in the inverted list, the spatial index of the respective posting list is searched to find those documents that intersect the query footprint. The text primary scheme was found by Vaid *et al.*, 2005 to have superior time performance to space primary. This was echoed by Zhou *et al.*, 2005 who also again demonstrated the merits of using R-trees rather than a regular grid with this approach.

Term	List of Document Ids.
Term 1	SpatialIndexT1[D1, D3, D6, D8]
Term 2	SpatialIndexT2[D2, D3, D7]
Term 3	SpatialIndexT3[D1, D2, D7, D9]
Term n	SpatialIndexTn[.....]

Figure 5.5: Text first spatio-textual indexing. Each text term in the dictionary is associated with a spatial index of the documents that contain. The term-specific spatial index is used to retrieve those documents that both contain the term and intersect the query footprint

Chen *et al.*, 2006 described a variation of the text primary method, the "Space Filling Inverted Index", in which spatial indexing was achieved by storing, for each text term, references to toeprints (see above), with the posting lists being sorted in toeprint ID order and hence in spatial order (as determined by the space filling curve). Retrieved toeprints are converted to document IDs using a table as in their Tile Index method (see section 5.5). Chen *et al.*, 2006 found this to have the best time performance of the methods that they implemented.

A variation of the latter approach is described in Christoforaki *et al.*, 2011 in which document IDs are based on space-filling curve spatial order as determined by the Z-order (which is equivalent to the sequence of Morton numbers), i.e. similar to toeprint IDs and, importantly, in which storage of the inverted index is optimised with the OPT-PFD algorithm (Yan *et al.*, 2009). To search the posting lists, a quadtree is used to determine the m ranges of document IDs for those documents that intersect the query footprint. If the inverted list is in main memory these ranges of docIDs are compared with the query footprint. If it is on disk, to avoid excessive disk accesses to fetch the contents of the posting lists, ranges of docIDs are merged into several (between 1 and 3) "sweeps" before being accessed. The contained m ranges are then searched in main memory as before. This approach is referred to as SFC-QUAD. In a variation of this method, SFC-SKIP, blocks of postings in the inverted lists are associated with minimum bounding rectangles to assist in filtering against the query footprint. SFC-QUAD, and a simpler method with inverted list optimisation but without spatial ordering of the document IDs, outperform in query time and storage costs, spatial-primary methods based on R*-trees with auxiliary information

(such as signatures, as described above) that filter the search to avoid following branches of the tree that lead to documents that do not contain the textual query terms. The method is designed however to retrieve all documents, rather than the top-k, that meet the spatial and text constraints of a query.

Inverted linear quadtrees

In the text-primary methods described above, the posting lists for each text term are organised in the form of either a regular grid index (Vaid *et al.*, 2005), an R-tree (Zhou *et al.*, 2005), or as a space filling-curve ordered list of the documents (Christoforaki *et al.*, 2011). A natural variation on the regular grid-based indexing approach is to take advantage of the superior performance of linear quadtree indexing in which, like some regular grids, location codes representing spatial cells serve as the keys in a database indexing structure such as a B^+-tree. This is performed in Zhang *et al.*, 2013 in which, for each text term, a linear quadtree is used to access each document that contains the term. Unlike some purely spatial quadtree indexes, here a quadtree structure is maintained explicitly in that empty leaf nodes are kept. Also each cell is associated with a signature to indicate whether the cell references an object (document) that contains the respective text term. When searching multiple linear quadtrees, i.e. one for each query term, this results in improved pruning of the search space. As each quadtree refers to a single text term, the signature can be represented by a single bit (rather than the lists of terms employed in the previous examples of a signature). When a quadtree cell is found that intersects the query footprint each document referenced by the cell must then be tested to check whether it is actually inside the footprint. Their implementation is designed to support top-k search and in an experimental comparison is shown to have superior query time to the CM-DIR-tree of Wu *et al.*, 2012, as well as lower storage costs.

Adapting to term frequency

A hybrid form of text-primary scheme, called S2I is described by Rocha-Junior *et al.*, 2011 in which a distinction is made between frequently and less frequently occurring terms, to support top-k queries. For each term that occurs in many documents, an R-tree is constructed with data stored in the non-leaf nodes relating to the maximum term frequency in the documents referenced by the respective subtree. This is called an aggregated R-tree (aR-tree) and is designed to support pruning of the search space according to decreasing relevance of the documents (referred to as "impact"). Documents containing less frequently occurring terms are stored in blocks that are analogous to inverted lists in that they store the identities of documents that contain a specific term. The entries (postings) in the block also contain the location (document footprint) of the document and the normalised term

frequency. When processing a query that accesses a block, the block data are ordered by spatio-textual relevance to the query prior to retrieving the top-k documents. If an aR-tree is accessed its non-leaf node data about term frequencies supports traversal in decreasing order of relevance to the query.

In an experimental comparative analysis of some spatio-textual access methods (Chen *et al.*, 2013) found that for AND queries that retrieve all documents that contain the query terms and occur in the query spatial footprint, the SFC-QUAD (Christoforaki *et al.*, 2011) method is the best (with regard to time and space). For top-k queries that rank the results by spatial and term relevance, the S2I scheme (Rocha-Junior *et al.*, 2011) was found to be best provided the number of query terms was fewer than five and that storage cost was not a major consideration. For queries with more than five terms the CDIR method was superior for query time. The study did not consider the Inverted Linear Quadtree (Zhang *et al.*, 2013) nor the SKIF-P scheme described below (though it does consider SKIF which treats document footprints as regions rather than points).

Inverted lists with spatial and spatio-textual keys

Very close integration of textual and spatial indexing can be obtained by treating location codes, representing the extent of spatial cells, in an identical manner to text terms within an inverted list, as proposed in (Vaid *et al.*, 2005) and implemented in (Khodaei *et al.*, 2012; Khodaei *et al.*, 2010). Thus the dictionary of the inverted index can contain both text terms, the posting lists of which refer to documents containing the respective term, and spatial terms, for which the posting list contains documents that relate to the region of space denoted by the location code or cellID. Processing a query requires converting the spatial aspects of the query expression to a set of location codes, as is required for the regular grid and linear quadtree schemes described earlier. These location code query terms are then added to the text query terms before accessing the index.

Even closer integration can be obtained by creating spatio-textual keys that concatenate text terms with cellIDs, so that now the posting lists refer only to documents that relate both to the respective text term and to the region of space given by the cellID. The query expression is then converted to a set of spatio-textual keys in which each text term is concatenated with those cellIDs that cover the query footprint. Spatio-textual key indexing was introduced in (Jones *et al.*, 2004) and shown experimentally to be clearly superior in query time to a scheme that separated text index and spatial index.

The spatio-textual key approach could be expected to lead to index size overheads due to creating keys for all actual combinations of text term and cellID, though Jones *et al.*, 2004 found the practical effect to be very much less than the theoretical worst case. The inverted linear quadtree scheme of Zhang *et al.*, 2013 can be regarded as a variation of the approach in that the linear quadtree of each term used in (Zhang *et al.*, 2013) is equivalent

to a set of inverted file records in which the keys concatenate a given text term with each quadtree cell that it occurs in. The difference is that in (Zhang *et al.*, 2013) the text term of each record of a linear quadtree is implicit, as there is only one text term per linear quadtree.

The indexing method of SKIF-P (Khodaei *et al.*, 2012) and (Khodaei *et al.*, 2010) describe an inverted index scheme in which the key may be either a text term or a cellID (i.e., locations code). The cellIDs represent spatial cells, though documents are represented by a single point footprint (but in their earlier version of the technique (Khodaei *et al.*, 2010) documents are represented by regions). They also integrate both spatial and textual relevance ranking by storing data for *spatial* $tf - idf$ that relates to the distance between the document footprint and the cellID of the dictionary, in addition to conventional textual $tf - idf$. SKIF-P was compared experimentally with the MIR-tree and CDIR tree (optimised versions of the IR^2-tree (De Felipe *et al.*, 2008) and IR-tree (Cong *et al.*, 2009) respectively). SKIF-P had superior performance in reducing the search time and amount of data accessed, i.e. when increasing the number of key words in the query (when greater than one); increasing the number of results k requested (as in top-k); and when modifying the relative weight of spatial and textual terms in the relevance calculation. However, it is not clear how SKIF-P compares in query time with the CM-DIR-tree or the Inverted Linear Quadtree (Zhang *et al.*, 2013).

5.6 Indexing in the exemplar systems

The accounts of the exemplar systems vary considerably in the extent to which efficient methods of spatial and textual indexing are used and in the extent to which the methods used are explained. Larson, 1996 focuses particularly on the GIPSY geographical information retrieval system (Woodruff and Plaunt, 1994), one of the earliest to extract and index geographic references from text documents and to discuss the associated challenges. Document footprints with geographically varying levels of confidence were estimated by combining polygonal footprints relating to individual place references. The authors refer to the overheads of processing the polygonal footprint data and the associated 3D representations of each document (with the third dimension relating to confidence in the relevance of the respective location), reflecting the fact that no efficient method for spatial data access was employed. It was suggested that a grid-based method for managing the data would have been advantageous.

The SPIRIT system (Purves *et al.*, 2007) was the basis for experimentation with alternative spatio-textual indexing methods in which the spatial indexing was based on a regular grid. The method that was selected as providing the fastest response time was an integrated text-primary system in which spatial indexes are used to manage the contents

of the postings lists of each text term. This is the only one of the exemplar systems that employs data access methods that provide close integration of textual and spatial indexing, as opposed to integrating the results of separate textual indexing and spatial indexing schemes. The STEWARD system (Lieberman *et al.*, 2007) employs a separate spatial index of document footprints, using a linear quadtree in combination with an inverted file of text documents. The order in which each index is processed is adapted to whichever of the spatial and textual component of the query has the apparently smaller search space. The NewsStand system (Teitler *et al.*, 2008) extends the STEWARD system, particularly with regard to the fact that related news items are clustered. It maintains an inverted index to manage the association between text terms and the clusters of documents in which they occur. Importantly, the NewsStand system is also based around a streaming concept which makes rapid updating of the index an important issue. This is in contrast to the other systems described here, which all describe more or less static applications of corpora.

The Virtual Itineraries in Pyrenees (PIV) system (Gaio *et al.*, 2008) also maintains separate indexes for textual and spatial data processing. It is notable for indexing not just the footprints of toponyms but the regions of space implied by relative expressions such as *east of X*. Minimum bounding rectangles (MBR) are derived to represent the regions and associate them in the index with the document paragraphs to which they refer. Query expressions are treated the same way to derive a query footprint in the form of an MBR. The indexed data are represented in an XML format and accessed with the XQuery language that performs matching of query and document footprints. There is no reference to conventional spatial indexing methods. Derungs and Purves, 2014 index the spatial occurrence of geo-referenced documents in their corpus with an adaptive grid similar to a quadtree in that cells can be divided into four. They start with a fixed grid of cell size 20km and then split individual cells into four quadrants if they reference more than a threshold number of documents, with a limit of four levels of subdivision (i.e., the smallest cell size is 5km).

The FrankenPlace exploratory search system (Adams *et al.*, 2015) implements a spatial index of text terms that consists of a multi-resolution hierarchy of triangular grid cells, in which individual cells subdivide into four triangles at each level. Terms, and hence their parent document or document sections, are associated with weights that are a function of the size and number of grid cells to which they relate. Each set of terms in a cell is referred to as a grid document. An inverted index of grid documents connects a term to its associated grid cells. There is also an index of documents that links documents to grid cells. Brown *et al.*, 2012 also present a graphical exploratory interface, in this case providing access to documents that have been georeferenced with the TextGrounder system and allocated to the cells of a latitude / longitude aligned grid, but they provide little detail on the indexing methods, the emphasis being upon the language model approach for assigning documents to locations. Wang and Stewart, 2015 use ArcGIS to manage the spatiotemporal references

to tornado events extracted from news articles. They do not refer specifically to the issues of spatio-textual indexing. The purpose of the Web-a-Where system (Amitay *et al.*, 2004) is toponym detection and resolution, rather than being an information retrieval system. As such the issue of spatio-textual indexing is not addressed.

5.7 Summary

Spatio-textual indexing methods are adapted to support queries that include both spatial and textual terms. They work by combining the characteristics of well-established indexing methods developed independently for spatial indexing and text indexing respectively. Spatial indexes are typically based on spatial keys that refer to spatial cells, where records for each cell list the stored objects that are either inside or intersecting the cells. Text indexing methods are usually based on inverted file (or inverted index) methods in which each text term in a dictionary of all unique terms in all documents in the database is associated with a "postings list" of the documents that contain the term. While spatio-textual indexing methods can use these independent indexes, most methods integrate them with either a space-first or text-first approach. In space-first indexes cells of the spatial index are associated with inverted text indexes that record just those text terms that are found in the documents associated with the cell. In text-first approaches, entries in each posting list can be organised in the form of a spatial index, or the dictionary of text terms can include spatial cell identifiers. Some of the most advanced spatial-first methods maintain data on term frequencies to support top-k queries to retrieve the potentially most relevant documents. A comparative study found that a text-first method with top-k query support was superior to a leading space-first approach for query time for queries with k less than 5. It may be noted that some of the text-first approaches are closer in structure to the inverted file based approaches assumed to be used in commercial search engine indexing methods and hence might lend themselves to indexing of heavily text based document retrieval (as opposed to some of the less text-heavy documents of some social media). As pointed out by Cong and Jensen, 2016 there is a need to further develop indexing structures which also deal effectively with real-time, streamed corpora (as opposed to the static collections mostly described in this article), which in turn emphasises the importance of future work indexing with respect to not only space and theme, but time. This emphasises perhaps the key research challenge with respect to indexing, considering exactly what is indexed (for example the nature of the geometry associated with a document or part of a document) and the relationships between the spatial (and temporal) and thematic component stored in the index.

6 Relevance Ranking

6.1 Introduction

This chapter discusses aspects of relevance central to GIR systems (Sections 6.2 & 6.3) along with spatial similarity and ranking measures (Section 6.4) and approaches for combining text and spatial similarity scores to produce a single ranked list (Section 6.5), ideally of relevant documents. Additionally, techniques to dynamically rank documents are being increasingly used (Section 6.6), as are methods to diversify results with respect to topic and location (Section 6.7). We end by discussing how relevance ranking is achieved in our exemplar systems.

6.2 Relevance and GIR systems

Given a query, GIR systems must retrieve and rank documents from the index that are *relevant* to the user's information needs. Saracevic, 1996 defines relevance in IR as "a criterion for assessing effectiveness in retrieval of information, or objects potentially conveying information". However as Case and Given, 2016 point out, the term relevance can have many meanings. In general, the term is used to denote something of interest. In information retrieval it typically refers to the connection between a query and results. A document is relevant if it is on the appropriate topic. However, from a user's perspective relevance will likely refer to how useful a document is in fulfilling a particular (often implicit) information need. This is commonly referred to as *pertinence* (or situational relevance) and deals with the situational and psychological perspectives of relevance: "Relevance is the property that assigns an answer to a question and pertinence is the property that assigns an answer to an information need" (Case and Given, 2016) (p.46).

Documents are typically ranked according to a relevance ranking function/model and top ranked documents are returned to the user. In standard (text-based) IR the relevance ranking function is based, at its simplest level, on computing the similarity of word matches between query and document, though in practice a wide range of features are used to attempt to encode important contextual information with respect to an individual user's information need. The main difference between standard IR and GIR is the central importance of the geographic relevance of documents to capture geographic proximity, containment and other spatial relationships. This is often achieved using spatial similarity and ranking methods. Most GIR systems handle the text and spatial components of search separately, combining the text and spatial similarity scores to produce a single ranked list.

Most GIR architectures include a *ranking* component. The ranker applies a relevance ranking function/model, $f(Q, D)$, to score the retrieved documents, each document denoted D, for a given query, Q. Documents are sorted according to their scores. In GIR the ranker

typically combines the spatial (where) and textual (what) attributes and a typical ranking process in GIR consists of the following steps (based on Martins *et al.*, 2005):

1. Transform the location and spatial operators (e.g., 'north of') in the query into one (or multiple) geometric footprints. Spatial operators could be implemented by modifying the geometry footprint associated with the toponyms in the query or by generating additional toponyms in the query (i.e., query expansion).

2. Quantify the degree of match between the query and document footprint(s) using a measure of spatial similarity.

3. Produce a ranking of documents matching the query footprint(s). Ranking is based on the similarity between the query and document footprints and combined with document rankings based on non-spatial similarity.

6.3 Notions of relevance for GIR

Similar to IR more generally, the notion of relevance must also be considered for GIR since a key part of a GIR system is to assign a relevance score to documents to estimate how well they are likely to fulfil a user's spatial information need. Raper, 2007 (p. 836) defines geographic relevance (GR) as "a relation between a geographic information need and the spatio-temporal expression of the geographic information objects needed to satisfy it" and argues it is particularly important for understanding the information seeking behaviours of mobile users. This is emphasised by the notion of a geographic information object rather than a document, where in Raper's model such objects are real-world entities which can be visited and utilised (for example a shop and its associated opening hours).

De Sabbata and Reichenbacher, 2012 describe GR more broadly, also encompassing users engaging with objects in the real world, for example in the case of mobile search: "GR refers to the relevance of a geographic entity, given a specific context of usage." In this usage a geographic entity again refers to individual physical entities in the real world rather than geo-referenced documents. The user's personal mobility and the geography of the environment are two distinguishing geographic features of GR in this setting. The former will involve criteria such as spatio-temporal proximity and directionality; the latter those such as the co-location of geographic entities. Although both of these frameworks focus on definitions of relevance defined with respect to representations of physical real world objects rather than documents, little other work has specifically focused on definitions of relevance in GIR, and most authors appear to simply transfer conventional models of relevance to the geographical context.

GIR systems (and location based services) thus aim to ensure that retrieved information is relevant with respect to the subject (topical or thematic) and also the spatial part of

Table 6.1: Criteria for assessing geographical relevance grouped into four sets (De Sabbata and Reichenbacher, 2012)

Properties	**Geography**	**Information**	**Presentation**
Topicality	Spatial proximity	Specificity	Accessibility
Appropriateness	Temporal proximity	Availability	Clarity
Coverage	Spatio-temporal prox.	Accuracy	Tangibility
Novelty	Directionality	Currency	Dynamism
	Visibility	Reliability	Presentation quality
	Hierarchy	Verification	
	Cluster	Affectiveness	
	Co-location	Curiosity	
	Association rule	Familiarity	
		Variety	

the user's query or their spatial location (e.g., current position). Relevance is judged based upon spatial relationships (e.g., overlap and adjacency) between the location expressed in the information need and spatial footprints identified within a document (Cai, 2002). The spatial and thematic aspects form basic conditions for relevance in a GIR system and coincide with the notion of topical relevance used in standard IR. In many situations, particularly those involving mobile users, time is also an important factor that influences relevance. For example, Palacio *et al.*, 2010 describe the PIV prototype GIR system that indexes and retrieves documents based on three dimensions: topical, spatial and temporal.

Counter to these relatively simple approaches to modelling relevance in GIR, De Sabbata and Reichenbacher, 2012 investigated 29 possible criteria related to geographical relevance for geographical objects. The criteria were grouped into 4 classes: those related to properties of the geographical entity (e.g., topicality and novelty); those related to geography (e.g., spatial proximity and directionality); those that capture how well the entity is represented in the information system (e.g., accuracy, reliability and specificity), and those used to judge how well the information is presented to the user (e.g., presentation quality and accessibility). The four sets of GR criteria are shown in Table 6.1.

The studies carried out aimed to examine users' assessments of the criteria in different contexts: in the first study participants were asked to imagine a situation where they have to find a geographic entity (i.e., place) in an urban environment and then rate statements representing the criteria. This provided a general opinion on the importance of different criteria, but lacked a specific task or context. In the second study participants were given specific scenarios and pre-defined maps with varying amounts of information about specific entities. Overall findings suggested that topicality and spatio-temporal proximity are fundamental criteria that are likely core to most, if not all, GIR tasks in the context of identifying relevant physical entities.

6.4 Computing spatial similarity

Measures of spatial similarity are used to estimate or infer the geographical relevance of documents to a particular query (see, e.g. (Hill, 2006; Larson, 2011; Cai, 2011)). The notion of geographical relevance can be argued to correspond to the relevance criteria listed under the 'Geography' column in Table 6.1. Query and document footprints can be compared to identify matching footprints; not necessarily as exact matches, but those with some degree of overlap between footprints that indicate they share some area in common. As well as simple overlap, *topological* relationships between a query and document footprint can be used to explore relevance for example in the form of *contains*, *is contained within*, *adjacency* and *overlaps*. *Geometric* relationships, based on distance and direction, may also be used, e.g., "20km north of". Standard GIS methods, such as those discussed earlier can be used here, though it is important to remember that simple spatial queries may not adequately represent the ways in which such concepts are conceptualised.

The ranking process in GIR is therefore based on quantifying the similarity between the query and document footprints. The similarity score is an estimate of relevance (or utility) to the user's information need. Documents retrieved are ranked in descending order of the scores. A common assumption is that documents that are spatially nearer to the query location will be more relevant to the user than those further away[1].

Commonly, GIR systems begin by identifying topological and geometric matches. A measure of spatial similarity is then used to compute the strength of the match based on the degree (or extent) of spatial overlap. Various measures can be used that may also take into account the relative sizes of the query and document footprints. This is similar to normalisation for document length in standard IR. For example, assuming Q is the area of the query region, D is the area of the document region, and O is the area of overlap between Q and D, the similarity, *sim(Q,D)* could be expressed as (Hill, 2006; Larson, 2011):

$$sim(Q, D) = 2 \cdot \frac{O}{(Q + D)}$$

This measure produces scores in the range 0 to 1 with 0 representing no overlap and 1 exact overlap between the query and document footprints. Larson and Frontiera, 2004 and Frontiera *et al.*, 2008 describe various other measures of spatial similarity, including the commonly used *Hausdorff distance*, a shape comparison metric that measures how far two subsets of a metric space are from each other. Simply for every point in Q the shortest distance to the other set D is found. The maximum distance found among all points in Q is kept as the Hausdorff distance.

Other factors might also be used to influence the similarity scores to prefer some documents over others, e.g., population counts. Measures of spatial similarity will vary

[1] This relates to Tobler's First Law of Geography (Tobler, 1970): everything is related to everything else, but near things are more related than distant things.

depending on footprint representation. For example, using Minimum Bounding Boxes or Rectangles the area of overlap could be computed. If document footprints were points, however, then *point in polygon* matching would be used to identify points contained within the query footprint. Other methods can be used to compute spatial similarity, such as distances based on information about ontological relations (Andrade and Silva, 2006); while combinations of relations, based for example on topological and distance methods, can also be used to compute a single spatial similarity score (Larson and Frontiera, 2004; Zaila and Montesi, 2015).

6.5 Combining thematic and spatial similarity

In GIR systems that handle the thematic and spatial components of a query separately there is a need to combine relevance scores to compute a single ranked list. This is because generally it is preferable that users browse a single ranked list of results. This can be achieved by computing a combined textual and spatial similarity score for each document (*score-based*) or by fusing ranked lists[2]. In the case of combining scores a common method is to use a linear combination (Andrade and Silva, 2006; Larson, 2011; Hariharan *et al.*, 2007b; Chen *et al.*, 2013). For a query, Q, and document, D, the combined score, *comb(Q,D)* is given by:

$$comb(Q, D) = \alpha_1 \cdot textsim(Q_t, D_t) + \alpha_2 \cdot geosim(Q_s, D_s)$$

where $textsim(Q_t, D_t)$ represents thematic similarity computed using, for example, the vector space model with BM25 term weighting (Robertson and Zaragoza, 2009), and $geosim(Q_s, D_s)$ spatial similarity. The weights α_1 and α_2 are weights that can be adjusted to reflect the relative importance of the geographic and textual components and $\alpha_1 + \alpha_2 = 1$. The scores for $textsim(Q_t, D_t)$ and $geosim(Q_s, D_s)$ are usually normalised within the range $[0, 1]$. Martins *et al.*, 2005 discuss other methods for combining scores, including various linear combinations.

Other approaches to combine ranked lists of documents involve using the rank positions of documents (e.g., Round-Robin and Borda Count), or using a combination of both score-based and rank-based methods (Palacio *et al.*, 2010). In the case of Borda Counts, each ranked list 'votes' for a document and the sum of the ranks from all ranked lists determines the score for each document from which a final ranking can be derived. Other approaches have involved probabilistic models to combine scores and predict the relevance of documents given a query (Frontiera *et al.*, 2008; De Sabbata and Reichenbacher, 2010). Finally, it is possible to use ranking measures which also emphasise the importance of diversity in a set of ranked documents; such approaches are briefly discussed in more detail below.

[2]The combination of multiple forms of evidence is often termed data-fusion or information-fusion

6.6 Learning to rank for GIR

In traditional retrieval the ranking function is created without training. Learning to Rank (LTR or L2R) is a commonly used approach for applying machine learning methods to automatically learn ranking functions from training data (Liu, 2009; Li, 2011). L2R is a supervised learning task that consists of training and test phases. Input data for the task are queries and document pairs; the output is a document ranking. Each query-document pair is represented by a *feature vector* where the features (or signals) are attributes of the query, document or query-document relationship, for example BM25, *tf–idf* or PageRank scores (Manning *et al.*, 2008 provide a good introduction to these common scoring techniques). Each query-document pair is given a label (relevance score or rank). The L2R algorithm trains a ranking model on the query-document pair feature vectors and labels that include relevance judgement information (e.g., relevance manually assigned or via click-through data). In the test phase, L2R will use the ranking model to predict labels from the feature vectors, i.e., produce a ranking of documents adapted to individual queries. The fitness of the predicted ranking from the correct ordering is described by defining a loss function, such as Mean Average Precision (MAP) or normalised Discounted Cummulative Gain (nDCG). Crucially, such approaches therefore require a set of labelled data – in commercial search such data are available in the form of query logs and click through data, but in research this is rarely the case.

In the case of GIR feature vectors can include thematic and spatial information, such as BM25 and spatial similarity scores. L2R can be used to alter the importance assigned to the spatial and thematic (and any other) components on a per-query basis and thereby provide a dynamic document ranking scheme. Martins and Calado, 2010 discuss learning to rank for GIR. They made use of state-of-the-art supervised machine learning methods, support vector machines (SVMs), to combine scores that will likely rank relevant documents nearer the top of the ranked list. This is achieved through the use of training data (i.e., GIR benchmarks) in which relevant documents are identified for queries and can be used to optimise for MAP. Based on 8 different ranking approaches, the authors show that the L2R method outperforms previous approaches based on heuristic combinations of features.

Shaw *et al.*, 2013 use machine learning methods to map an estimate of a user's current location to a semantically meaningful place of interest, e.g. home, restaurant, or store. A spatial search algorithm is developed that infers a user's location by combining aggregate signals mined from billions of Foursquare check-ins with real-time contextual information. Machine learning methods are used for predicting user models and ranking local search results. These outperform common methods for location search based on distance and popularity. Note that in this context (that of physical entities visited by users) much more potential training data, in the form of user generated content (such as FourSquare check-ins, Tweets, etc.), exists than is the case for document search.

L2R provides the opportunity to integrate further spatial and document features into the ranking model that may better help to capture the relevance criteria as shown in Table 6.1. However, a major challenge remains the limited availability of training data outside of commercial search contexts. Ferrès and Rodriguez, 2015 demonstrate the possibilities of using existing training data, in their case from the GeoCLEF evaluation campaign discussed in detail in Chapter 7 on evaluation, to improve over results through more intelligent ranking.

6.7 Diversity in GIR

As well as ranking documents, users are often also interested in having an appropriately diverse set of relevant documents. Thus, when presented with a set of results users may prefer a range of relevant documents, for example different media types or subjects, rather than a list containing repeated references to very similar documents. This has led to a field of research in IR called *diversification* in which the goal is to select documents that are relevant but also different from each other (Santos *et al.*, 2015; Carbonell and Goldstein, 1998; Zhai *et al.*, 2003; Drosou and Pitoura, 2010).

In the case of GIR some research into *spatial diversity* has been carried out in which the IR system attempts to provide results that are both relevant and also spatially diversified, i.e., from many different (but spatially relevant) locations (Kreveld *et al.*, 2005; Tang and Sanderson, 2010). Tang and Sanderson, 2010 describe spatial diversity as the "more locations that are covered and more intense the coverage is, the better spatial diversity a list of documents achieves, which may satisfy users better." For example, consider the query "castles in UK". Relevant documents are those describing castles that are located in the UK. However, users may prefer results that include castles located throughout the UK rather than clustered in a single place. This intuition was confirmed in the experiments by Tang and Sanderson, 2010. They perform an experiment to compare results with and without spatial diversity applied using Amazon's Mechanical Turk and crowd workers and show that in general users prefer results that are spatially diverse over those that are not. Results from a GIR system are re-ranked using a diversity score that gives higher weight to a document that is not only relevant but also far away from previous ones encountered in the results list. Algorithms based on *spatial coverage* (a measure of location coverage) are used to ensure as many different locations are covered within the higher-ranking results as possible.

6.8 Relevance ranking in the exemplar systems

In the following we provide descriptions of three contrasting approaches to ranking taken by a selection of our exemplar systems.

In the description of STEWARD, Lieberman *et al.*, 2007 state how documents are ranked with respect to queries in three scenarios: queries with a purely geographical component, a keyword components and a combination of both. In the first scenario, documents are ranked by the extent to which the geographic entity in the query serves as the focus of the document. Features used to determine this include the number of times proximate geographic locations are mentioned in the document and their distribution throughout. In the second scenario, when the query consists of non-graphical keywords, STEWARD ranks documents according to frequency and distribution of keywords throughout the document. In the third case, documents containing keywords (i.e., topically relevant) are ranked in increasing order of distance from their geographic focus to the query.

In the case of the SPIRIT system (Purves *et al.*, 2007) documents are ranked according to both textual and spatial relevance. From a geographical perspective, each document is represented by a "bag of footprints" derived after recognition and resolution of all toponyms found in a document. The query is also represented as a footprint. Unlike many of the exemplar GIR systems, SPIRIT is able to handle spatial operators in the query, such as 'inside' and 'near'. Depending on the spatial relationship used, different formulae are used to calculate footprint similarity scores between query and document footprints. When all footprints in a document are assigned a similarity score with respect to the query footprint, a document spatial similarity score for the document can be calculated. The relevance ranking component combines the spatial and textual document scores (based on BM25 term weighting) to generate a single ranking. Documents are ranked in descending order of their document similarity scores. Various methods for combining evidence from spatial and textual similarity are used to rank the retrieved documents (Kreveld *et al.*, 2005).

Both of the above approaches start from the premise that documents should, at least in some cases, be presented to the user as a ranked list. The NewsStand system (Teitler *et al.*, 2008) takes a quite different perspective, and aims to "group articles into story clusters based on their textual and geographic content." Here the key task is thus to cluster articles, firstly according to content (and because news is an essentially temporal phenomena) and time using Cosine similarity with a Gaussian attenuator taking account of the temporal distance of a document from the cluster. By processing the locations associated with news stories found in cluster, it is possible to assign a cluster focus which is then used to display information directly on a map. This approach is much more similar to the visual information seeking mantra proposed by Shneiderman: "overview first, zoom and filter, then details-on-demand" (Shneiderman, 1996) where, crucially, the underlying assumption is that the user

be allowed to display a, suitably summarised, overview of the complete set of results as an initial stage. This directly links the ranking process to the user interface (c.f. Chapter 3).

6.9 Summary

In this chapter we have discussed the concept of relevance ranking: the ordering of documents returned from a search system by some notion of relevance, with the goal of ranking more relevant documents higher in the results list. In the case of GIR this involves dealing with various measures of similarity computed between the query and documents as represented in the index and commonly combining such measures to allow a single ranking of results. Various approaches exist for this with more recent attention being paid to machine learning methods in the form of learning to rank. However, the effective (and efficient) ranking of documents is still an area of active research interest, especially as most systems make use of simplistic notions of relevance and with limited use of contextual information that is necessary to inform a more situational view of relevance. Furthermore, the lack of available training data has limited the development of approaches which take advantage of methods based around learning to rank, especially where geographic relevance is considered. Furthermore, though initial research has demonstrated the potential advantages of spatial diversity for ranking in GIR, again little work has followed up on this notion.

7 Evaluation

7.1 Introduction

In this chapter we discuss approaches for evaluating GIR systems. We begin by discussing the need for evaluation(Section 7.3) , before exploring briefly evaluation in IR generally (Section 7.3) and then go on to describe common evaluation methods and metrics for quantifying aspects of search quality within GIR (Section 7.4). We end by discussing evaluation approaches used in two of our exemplar systems (Section 7.5).

7.2 The need for evaluation

In previous chapters we discussed topics central to developing the necessary components of a GIR system. Along the way there are many questions to address and decisions to make. For example, which particular indexing or georeferencing method should we use? Should we present search results in a map-based or list-based format? Which particular method works 'best' for combining similarity scores? A key remaining question is establishing how well the GIR system meets its goals of retrieving relevant documents and more widely in helping its

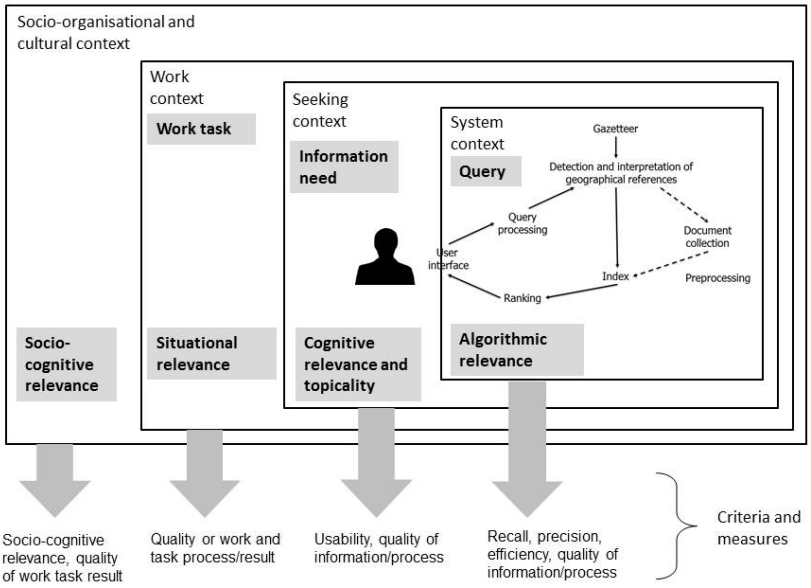


Figure 7.1: Layered contexts of a GIR system that can be considered for evaluation (based on Järvelin, 2011)

users to fulfill their information needs, making decisions and completing their search and work tasks. To address these questions, and more, highlights the need to evaluate: "to judge or calculate the quality, importance, amount, or value of something[1]". When preparing an evaluation, key questions include (Van Rijsbergen, 1979): (i) *why* conduct the evaluation (i.e., the goal/purpose of the evaluation), (ii) *what* should be evaluated (i.e., the success criteria and measures), and (ii) *how* the evaluation should be conducted (i.e., the evaluation methodology). It may also be helpful to identify the intended stakeholders in the evaluation. In this chapter we will consider these aspects and more. Evaluation activities are required to build effective, efficient and usable IR systems that enable aspects of the search process and its outcomes to be quantified and measured. This can be used for purposes such as benchmarking and monitoring performance, investigating the potential benefits of new features and to enable improvements to be made. We start by considering evaluation in IR more generally.

[1]Cambridge Dictionary: http://dictionary.cambridge.org/dictionary/english/evaluate

7.3 Evaluation in IR

Evaluation has been an active area of investigation for many decades in IR, particularly experimental evaluation (Harter and Hert, 1997). Initially more emphasis was placed on IR system performance (a system- or algorithmic-oriented view) and assessing the quality of search results through *retrieval effectiveness*. However, more recently there has been a focus on human-oriented perspectives (a user-oriented view), including on user experience, the search process and wider aspects of discovery and information use (White, 2016). Additionally, there has been increased interest in evaluation of operational IR systems and associated methods, for example *online evaluation*: "the evaluation of a fully functioning system based on implicit measurement of real users' experiences of the system in a natural usage environment" Hofmann *et al.*, 2016 (p. 3).

Harter and Hert, 1997 use the following definition to capture the notion of IR evaluation: "The process of identifying and collecting data about specific services or activities, establishing criteria by which their success can be assessed, and determining both the quality of the service or activity and the degree to which the service or activity accomplishes stated goals and objectives." This is a useful definition to reflect on as it includes the following: the notion of evaluation being a *process* or methodology (e.g., controlled experiment, such as a lab-based experiment, use of test collection or controlled online experimentation); the need to establish *criteria* by which success can be determined, which will involve the use of measures or metrics to quantify the criteria; and the need to define *goals* and objectives: typically goals of the system or some component under test but could also include the goals of the evaluation, such as research hypotheses (Vakkari and Huuskonen, 2012; Vakkari, 2012).

Multiple approaches are needed to evaluate IR systems in a holistic manner, both during development of the system or its sub-systems (*formative* evaluation) and at the end of development (*summative* evaluation). In the case of developing operational systems there may also be on-going monitoring and evaluation activities over time, typically through the process of developing appropriate so–called Key Performance Indicators (KPIs). Evaluation becomes increasingly involved as IR systems go beyond just providing a search box to offering a rich array of features (e.g., facets and categories, recommendations, visualisations, etc) to support the user's search and discovery discovery (Wilson, 2011). In such cases simply treating evaluation as a process of assessing query–results is not sufficient. However, assessing the quality of search results remains a dominant activity in IR evaluation.

7.3.1 Contexts of evaluation

IR systems are used by people to solve problems and perform tasks and sit within a broader context of use that should be considered for evaluation (Dunlop, 2000). Evaluation in IR

is often distinguished between the following levels (Saracevic, 1995; Järvelin, 2011): (1) evaluation within the IR system context; (2) evaluation within the information seeking context; (3) evaluation within the work context; and (4) evaluation within the socio-organisational and cultural context (see Figure 7.1). In one sense this depends on where one places the boundary of 'system' when referring to an 'IR system' (Robertson and Hancock-Beaulieu, 1992). Considering such levels helps to identify the scope of evaluation (e.g., to what extent users and the wider context are included in the system and therefore taken into account). This partly addresses the 'what' aspect of planning an evaluation activity. Figure 7.1 also indicates how the notions of relevance change at the various levels, which is important given that one of the primary goals of an IR system is to return information *relevant* to the user's information need. There has been continual debate about the relationship between the levels, especially for performance measures (e.g., to what extent does the quality of search results effect user satisfaction or predict task success) (Vakkari and Huuskonen, 2012; Al-Maskari *et al.*, 2007). The diagram also shows how evaluation criteria and measures varies across the levels of context.

7.3.2 System- vs. user-oriented approaches

Two major paradigms of evaluation in IR research exist (Järvelin, 2011): (i) systems-oriented (including IR algorithm development and evaluation, and limited human-system interaction); and (ii) user-oriented evaluation (including studies of human information behaviour, information seeking and interactive information retrieval). For decades, the primary approach to IR evaluation has been system-oriented (or batch-mode). The focus of system-oriented approaches is on retrieval algorithms and their outputs – the ability to discriminate relevant from non-relevant documents and to rank the results effectively. This might consider the quality, diversity and position of relevant documents in the ranked list[2]. One of the most widely used methodologies for conducting IR experimentation is referred to as the *Cranfield approach* or methodology (Cleverdon, 1991). This approach to evaluation makes use of *test collections* (also referred to as reference collections) that allow systematic and repeatable evaluations to be carried out in a controlled manner (Sanderson, 2010; Harman, 2011; Clough and Sanderson, 2013). The popularity of the Cranfield approach can partly be attributed to its use in the U.S. NIST-funded Text REtrieval Conference[3].

User-oriented evaluations, on the other hand, seek to measure how the system as a whole facilitates the process of a user seeking and retrieving information and people are involved

[2]Assessing relevance in GIR requires taking into account criteria beyond topic, for example space and time. This must be dealt with during relevance assessment and make use of notions of relevance as described in Chapter 6.

[3]The TREC series of large-scale evaluation campaigns utilising the Cranfield Model began in 1992 and have since stimulated significant developments in IR (Voorhees and Harman, 2005; Harman, 2011), see: http://trec.nist.gov/

somehow to assess the success of the IR system, or components of the system (Borlund, 2009; Kelly, 2009; White, 2016). In this view the success of searching depends not just on the IR system but on the combination of system and user (e.g., their search literacy skills, domain knowledge, etc.). Approaches for conducting user-oriented evaluation typically fall within the areas of *Interactive IR (IIR)* and *Human Computer Information Retrieval (HCIR)*, merging evaluation approaches from IR and Human Computer Interaction. Criteria used to assess retrieval systems are typically concerned with how well users achieve their goals or tasks, and their success and satisfaction with the search results. A common approach to user-oriented evaluation involves recruiting 'users' (real end users of the system or substitute volunteers) to participate in search tasks in a controlled lab-based environment (Hoeber and Yang, 2007; Kelly, 2009). Their interactions with the system are recorded, along with feedback on the system and information about their individual characteristics (e.g., age and cognitive abilities). Alternative types of user study include comparing results from multiple systems in a side-by-side manner (Thomas and Hawking, 2006), A/B testing, where a small proportion of traffic from an operational system is directed to an alternative version of the system and the resulting user interaction behaviour compared (Manning *et al.*, 2008) and use of search engine log data to observe how users' click patterns for relevant documents vary (a form of *implicit* user feedback) (Hofmann *et al.*, 2016).

7.3.3 Evaluation methodologies

A further key consideration for evaluation is the 'how' aspect, or the methods used for data collection. Needless to say many different methods exist and White, 2016 identifies nine commonly used approaches, broadly grouped into *naturalistic* (field studies, instrumented panels and log analyses) and *controlled methodologies* (interviews and focus groups, laboratory studies, crowdsourcing, surveys, online methods such as A/B testing, and offline methods such as the Cranfield Model and simulations). In practice the approaches used will vary depending on the focus of the evaluation (e.g., system or user), what data is needed (e.g., clicks on relevant documents, user satisfaction scores, etc.), whether evaluation is being undertaken on an operational system, etc. In summary three key methodologies are offline testing using the Cranfield Model and test collections (Sanderson, 2010; Harman, 2011), user testing (Borlund, 2009; Kelly, 2009; White, 2016) and online testing (Hofmann *et al.*, 2016).

Often evaluation, whether system-oriented, user-oriented or evaluating an operational system, will typically follow an experimental setup comprising *dependent variables* (e.g., performance measures) and *independent variables* (e.g., different systems/components under test). The experimental setup is used to attempt to isolate confounding variables to ensure that the differences in dependent variables are likely to have come from the independent variables (e.g., particular system/algorithm). In the case of systems-oriented evaluation

the test collection can provide a controlled experimental environment where the dependent variable (i.e., precision or other measures as described in Section 7.3.4) is measured based on search output, and independent variables typically include different systems, algorithm and parameters. In the case of a user-based design, dependent variables may include measures capturing the user's search process and strategy (e.g., duration of session, number of query reformulations, use of search/browse strategies, etc.), measures of the search output (e.g., precision, user satisfaction, etc.) or outcome variables (e.g., work task completion, perceived quality of outputs, etc). Independent variables could include different systems (e.g., user interface A vs. B) or components (e.g., new interface feature A vs. B) which in this case the user and task variables are fixed; or the searchers where their characteristics (e.g., knowledge, cognitive style) are varied and the search engine and interface features fixed. Often the inter-relations between variables are not well studied or understood (which is necessary in being able to explain the outcomes).

7.3.4 Criteria and measures

The evaluation must also consider the 'what' that reflects what evaluation criteria and measures are used. Cleverdon *et al.*, 1966 listed six criteria that could be used to evaluate an IR system: (1) coverage, (2) time lag, (3) recall, (4) precision, (5) output presentation and (6) user effort. This is a broad view of 'system' in the sense that it goes beyond the IR components (e.g., indexing, matching and ranking) to include the coverage of the collection with respect to a topic (coverage) and incorporates elements of the user search process (user effort) and system usability (output presentation). The most well known measures are *precision* and *recall* (see, e.g. (Manning *et al.*, 2008)), which capture the ability of an IR system to discriminate relevant from non-relevant documents (a *binary* relevance decision). Precision measures the proportion of retrieved documents that are relevant; recall measures the proportion of relevant documents retrieved. However, these measures do not take the ordering of results or characteristics of users' search behaviour into account. For example, research has shown that users are more likely to select documents higher up in the ranking (*rank bias*) and, unsurprisingly, start at the top of a ranked list and work their way down. To accommodate for this *ranked-based* evaluation measures are commonly used, such as Average Precision (AP) that measures precision at fixed levels of recall, and Mean Average Precision (MAP) where AP is averaged across multiple queries. Measures have also been developed that accommodate non-binary relevance assessments (*graded relevance*), such as Discounted Cumulative Gain (DCG) and its normalised version, normalised Discounted Cumulative Gain (nDCG) (Järvelin and Kekäläinen, 2000). Extensions to DCG have also been proposed such as Expected Reciprocal Rank (ERR) which incorporate the length of time a user will take to find a relevant document (Chapelle *et al.*, 2009).

Further examples of categories of evaluation criteria include Su, 2003, who identifies four main evaluation criteria for web search engines: (i) *relevance* (assessment of the quality of search results); (ii) *efficiency* (time, effort or costs involved in the retrieval process); (iii) *utility* (usefulness/value of the information retrieved to information needs/tasks); and (iv) *user satisfaction* (the degree with which a system meets the needs of its users or how much users like the system). This might include satisfaction with system features and interaction, user interface and ease of use, search results and overall performance, etc. More recently, White, 2016 in his book on interaction in search systems describes two main categories of performance measures: (i) those based on the search *process* (e.g., learning, user effort, cognitive load, enjoyment, frustration and engagement), and (ii) measures reflecting the search *outcomes* after the search process is complete (e.g., relevance-based metrics, novelty, diversity, search success and user satisfaction). In addition, it may also be common to evaluate other aspects of the system such as information quality and usability of the user interface. Established evaluation frameworks can be used, such as the HEART (Happiness, Engagement, Adoption, Retention and Task success) framework for assessing user experience (Rodden *et al.*, 2010) and the System Usability Scale (SUS) (Brooke, 1996) or the Computer System Usability Questionnaire (QUIS) (Lewis, 1995) for assessing usability.

There are literally hundreds of measures and metrics proposed for quantifying evaluation criteria. Kelly, 2009 helpfully groups measures into the following four categories: (i) *performance measures*: evaluation of the retrieval process or output (e.g., Precision, Recall, P10, GMAP, BPref, DCG, response time, informativeness, cost and utility measures etc.); (ii) *interaction measures*: these are used to assess the search process and typically reflect usage activities (e.g., number of queries, number of clicked items, etc.); (iii) *usability measures*: the extent to which users can use a system to complete a task in specific context of use (typically with criteria of effectiveness, efficiency and satisfaction); and (iv) *contextual measures*: typically these are aspects of the user (e.g. individual characteristics such as intelligence, creativity, personality, memory and cognitive style) or context of use (e.g., time, location, task difficulty etc.).

7.4 Evaluating GIR systems

The preceding material set out some basics of evaluation in information retrieval. it is important to emphasise that most efforts in evaluating GIR have followed similar approaches, driven by both the backgrounds of those working in GIR, and the publication cultures in which researchers operate, emphasising for example a need for test–collection based measures of retrieval performance.

Thus, similarly to distinguishing levels of content for IR evaluation (Figure 7.1), Mandl, 2011 describes four levels at which GIR systems can be evaluated: (i) at the *component level*

(i.e., evaluating individual parts of an IR system); (ii) at the *system level* (i.e., evaluating the outputs of a complete IR system using test collections); (iii) at the *user-system-interaction level* (i.e., assessing an entire GIR system including interfaces and visualisations in a controlled laboratory setting); and (iv) at the *user performance level* (i.e., assessing an operational system in use and its impact on daily work tasks).

Evaluation in GIR has typically focused on more system-oriented approaches and assessing search output, including the construction of standardised benchmarks for systems and individual components (Palacio *et al.*, 2010; Bucher *et al.*, 2005; Wallgrün *et al.*, 2017; Gritta *et al.*, 2017). However, it is important to evaluate GIR systems and their constituent components (e.g., for geo-referencing, ranking and presentation/visualisation) using a variety of approaches, assessing both system performance and users' interactions with and preferences for certain interfaces (Bucher *et al.*, 2005; Mandl, 2011). Particular aspects that must also be accommodated in evaluations are assessing spatial criteria, such as relevance, and evaluating visualisations commonly found in GIR interfaces, such as maps. More in-depth, task-based user-studies tend to be used where the focus is on the interaction between the user and the interface in areas such as geographic information visualisation (Opach *et al.*, 2013) or digital map interaction (Wilkening and Fabrikant, 2013). There is clearly a distinction between assessing the components of a GIR system, perhaps individually and during system development, compared to evaluating an integrated working system incorporating a user interface. Martins *et al.*, 2005 identify various aspects of GIR systems that can be evaluated: (i) building geographical ontologies to assist GIR; (ii) handling geographical references in text; (iii) assigning geographical scopes to the documents; (iv) ranking documents according to geographical relevance; and (v) building user interfaces for GIR. Typically, standard IR evaluation approaches are adopted for assessing the performance of GIR systems (Palacio *et al.*, 2010; Bucher *et al.*, 2005).

In the following sections we start by describing the assessment of relevance in GIR (Section 7.4.1), then we describe standard benchmarks that have been created to provide the resources needed to evaluate GIR systems (Section 7.4.2). Typically these have been constructed within the context of evaluation campaigns (e.g., GeoCLEF). Next, two case studies are discussed: the creation of a test collection for system-oriented evaluation of the PIV system (Palacio *et al.*, 2010), and evaluation of the SPIRIT prototype GIR system that, unusually for GIR, makes use of not only system-oriented but also user-oriented evaluation approaches (Bucher *et al.*, 2005). This highlights how various approaches can be used to evaluate GIR systems based on the IR evaluation methods previously discussed.

7.4.1 Assessing relevance

Robertson, 1981 argues that relevance should be treated as a continuous variable and researchers have experimented with judging relevance using various levels or scales. The

most popular scales for assessing relevance are binary and graded relevance scales. Judgments based upon a *binary* relevance scale are used to assess whether a document is relevant or non-relevant with respect to a given information need or query. In the case of making *graded* relevance a scale with multiple categories, such as *highly relevant*, *partially relevant* or *non-relevant*, is used. The additional benefit of using graded relevance scales is that a wider range of system effectiveness measures, such as *Discounted Cumulative Gain (DCG)*, can be used. Beyond relevance other criteria may be evaluated such as the usefulness, quality or authority of results, and users' satisfaction with the results (Mao *et al.*, 2016).

By way of example, in the evaluation of the SPIRIT system document relevance was measured with respect to the two dimensions of spatial and thematic relevance (Clough *et al.*, 2006; Purves *et al.*, 2007). The scheme was devised to gather relevance judgments for the purposes of system-oriented evaluation. The two dimensions (topicality and spatial) were assessed independently using a 3-point (graded) relevance scale with a simpler binary scheme eventually used to assess topicality:

- **Topical relevance:**
  - **Score 1** - means the document is relevant since it points to a resource with information about the query concept.
  - **Score 2** - means the document does not provide information about the query concept.
- **Spatial relevance:**
  - **Score 1** - means the document refers to a location that is in/near the query location AND you think that the location in the document has sufficient detail for you to find it on a local map of the area.
  - **Score 2** - means the document refers to a location that is in/near the query location BUT you think that there is insufficient information for you to find that location on a local map of the area.
  - **Score 3** - means the document does not fall within the query location.

Based on using the scheme, Clough *et al.*, 2006 showed that, in general, spatial relevance was more difficult to judge than thematic relevance. Purves and Clough, 2006 went further to demonstrate the effects of an assessor's geographic knowledge on making reliable and valid relevance assessments: judges with little knowledge about a query region were unable to reliably assess a document's spatial relevance. This result relates to different levels of cognition of geographic space, with more detailed knowledge tending towards so-called survey knowledge (a metric understanding of how objects are arranged in space) being necessary to complete such evaluation tasks (Mark *et al.*, 1995). This result can be related

to other work which has also demonstrated that domain expertise affects the quality of relevance assessments obtained (Bailey *et al.*, 2008; Kinney *et al.*, 2008). However, it is also worth emphasising the importance of an understanding of spatial cognition as a potential input to the GIR process (Montello, 2016).

Gathering relevance judgments is highly time-consuming and often creates a bottleneck during evaluation. Various techniques have been proposed for reducing this bottleneck including document sampling, the use of interactive search and judge techniques, and the simulation of queries and relevance judgments based on search logs (see (Clough and Sanderson, 2013)). A technique that has received considerable interest recently is the use of *crowdsourcing*, whereby the job of assessing relevance is outsourced to an undefined, generally large, group of people in the form of an open call. Amazon Mechanical Turk (AMT) is one such example of a crowdsourcing platform and has recently been used successfully to gather relevance assessments in IR (Alonso and Mizzaro, 2009; Carvalho *et al.*, 2011). De Sabbata *et al.*, 2012 also experimented with using crowdsouring to gather relevance assessments for a location-based service evaluation, which can be seen as analoguous in many ways to evaluation tasks typical in GIR. They compared results obtained using crowdsourcing with classical methods for gathering relevance and demonstrated the feasibility of the approach. In more recent work, Reichenbacher *et al.*, 2016 used crowdsourcing to explore the efficacy of different ranking methods which include dimensions of geographic relevance important in the problem of identifying physical entities (i.e., search in a location-based context). In addition to using crowdsourcing methods, the use of *implicit* feedback based on users' clicks on results can also be used as a proxy for relevance.

7.4.2 GIR evaluation resources

Palacio *et al.*, 2010 discuss evaluation resources for GIR, and in particular evaluation of systems that utilise and handle topical, spatial and temporal information. They highlight a number of GIR projects undertaken between 1994-2010 in which different evaluation resources have been developed and made available. A number of evaluation frameworks are also described, including those within TREC-style evaluation campaigns: GeoCLEF, GeoTime and GikiCLEF. These evaluation initiatives have addressed topical, spatial and temporal aspects and created, at least in theory, reusable evaluation resources. However, many of these resources have seen limited use beyond the initial campaigns or the authors of the original studies, despite a clear need for better evaluation resources for GIR. One main reason for this lack of reuse is likely related to the complicated licensing requirements involved in reusing the original data, and the related judgments.

GeoCLEF

The lack of standardised resources for evaluating GIR systems was recognised in the European Cross Language Evaluation Forum (CLEF) and led to the organisation of an ad-hoc geo-spatial retrieval task in CLEF 2005 (called GeoCLEF[4]) (Mandl *et al.*, 2008b; Cardoso, 2011). This enabled a TREC-style evaluation of cross-language GIR systems and the development of a benchmark for monolingual and cross-lingual GIR. The datasets used in GeoCLEF2005-08 consisted of news articles from various media sources and languages (see Table 7.1). The goal of GeoCLEF was not only to create standardised and reusable resources for evaluating GIR systems based on the Cranfield framework, but also to bring together researchers working on GIR to stimulate discussion.

Table 7.1: Collections used in GeoCLEF

Language	Sources (year)	#docs	Used at
English	Glasgow Herald (1995), Los Angeles Times (1994)	169,477	GeoCLEF2005-2008
German	Der Spiegel (1994-95), Frankfurter Rundschau (1994) Schweizer Depeschen Agentur (1994-95))	294,809	GeoCLEF2005-2008
Portuguese	Público (1994), Folha de São Paulo (1995)	210,734	GeoCLEF2006-2008
Spanish	Spanish Agency EFE (1994-95)	454,045	GeoCLEF2006

GeoCLEF is an example of system-oriented evaluation where the focus was on aspects, such as query translation, query expansion, translation of geographical references, use of text and spatial retrieval methods, retrieval models and indexing methods. In most years participants could provide monolingual submissions (*runs*) where both topics and documents are in the same language, or bilingual where the topics in language X are translated to match documents written in language Y (where X≠Y). Table 7.2 summarises the topic and document languages used in GeoCLEF 2005-08, together with the numbers of participating groups.

Across the 4 years of GeoCLEF the organisers provided 100 topics (25 per year) that were generated by hand and designed to be of a geo-spatial nature, including both locations and spatial relations. Topics included, for example, "Shark Attacks off Australia and California" (Topic 001) and "Cities within 100km of Frankfurt" (Topic 027), "Car bombings near Madrid" (Topic 030), "Malaria in the tropics" (Topic 034). Topics were structured similarly to TREC topics and consisted of a title, narrative and description. Two example GeoCLEF topics from the 2005 and 2006 editions respectively are shown in Table 7.3.

[4] http://www.uni-hildesheim.de/geoclef/

Table 7.2: Summary of topics and collections used in GeoCLEF and number of participating groups

Campaign	Collection languages	Topic languages	#groups (#runs)
GeoCLEF 2005	English, German	English, German	12 (117)
GeoCLEF 2006	English, German, Portuguese, Spanish	English, German, Portuguese, Spanish, Japanese	17 (149)
GeoCLEF 2007	English, German, Portuguese	English, German, Portuguese, Spanish, Indonesian	13 (108)
GeoCLEF 2008	English, German, Portuguese	English, German, Portuguese	11 (131)

Topics in GeoCLEF varied in their complexity with respect to the degree of reasoning required for a high retrieval score. This was required to demonstrate the benefits of incorporating spatial reasoning into the IR system beyond using keyword-based approaches only. This included incorporating the following attributes into GeoCLEF topics, particularly in 2006-08 (Mandl *et al.*, 2008b):

- Ambiguity, including places in different locations (e.g., St Paul's Cathedral in London and São Paulo) and different names for the same place (e.g., Greater Lisbon, Grande Lisboa ang Großraum Lissabon)
- Imprecise/vague geographic regions (e.g., Western Europe)
- Geographical relations beyond *in* (e.g., near, within n km of)
- Granularity below the level of country (e.g., fairs in Lower Saxony)
- Complex region shapes (e.g., along the rivers Danube and Rhine)

Table 7.4 shows the titles for the 25 topics used in the GeoCLEF 2006 campaign, together with the number of relevant documents. Mandl *et al.*, 2008b identify the best and worst performing topics from the 2006 dataset. The two worst performing topics were "Wine regions around rivers in Europe" (maximum AP=0.0172) and "Cities within 100km of Frankfurt" (maximum AP=0.0359), both of which require appropriate spatial modelling. The best two performing topics were "Fishing in Newfoundland and Greenland" (maximum AP=0.9161) and "Car bombings near Madrid" (maximum AP=0.7862).

The organisers of GeoCLEF provided relevance judgments (qrels) based on assessing pools of documents generated by participant submissions. In total over 100,000 relevance

Table 7.3: Example topics used in GeoCLEF

```
<top>
<num> GC023 </num>
<EN-title> Murders and violence in South-West Scotland </EN-title>
<EN-desc> Find articles on violent acts including murders
in the South West part of Scotland. </EN-desc>
<EN-narr> A relevant document will give details of either specific
acts of violence or death related to murder or information about
the general state of violence in South West Scotland. This includes
information about violence in places such as Ayr, Campeltown,
Douglas and Glasgow. </EN-narr>
<EN-concept> Murders and violence </EN-concept>
<EN-spatialrelation> South-West of </EN-spatialrelation>
<EN-location> Scotland </EN-location>
</top>
```

```
<top>
<num>GC030</num>
<EN-title>Car bombings near Madrid</EN-title>
<EN-desc>Documents about car bombings occurring near
Madrid</EN-desc>
<EN-narr>Relevant documents treat cases of car bombings
occurring in the capital of Spain and its outskirts
</EN-narr>
</top>
```

judgments were made on the topics. The resulting relevance assessments were used to compute Mean Average Precision (MAP), Recall and Precision at K. Across the 4 years of GeoCLEF 33 research groups participated in the activities and submitted 505 experiments (or runs). More than 100,000 relevance judgments were carried out to generate the evaluation resources (Mandl, 2011). Gey *et al.*, 2007 and Mandl, 2011 describe problems with evaluating GIR systems in the GeoCLEF framework. These include the difficulty in generating topics whereby the benefits of handling spatial aspects of queries could be observed compared to standard IR indexing and retrieval, the difficulty in gathering reliable relevance judgments, and problems with task participation resulting in difficulties with generating sufficient judgment pools. Nonetheless, despite these issues GeoCLEF was a landmark in attempting a community–based evaluation of GIR, and formed the basis for much ongoing research, especially in a European context.

Table 7.4: GeoCLEF 2006 topics (taken from (Andogah, 2011))

ID	Title	#relevant
GC026	Wine regions around rivers in Europe	9
GC027	Cities within 100km of Frankfurt	19
GC028	Snowstorms in North America	19
GC029	Diamond trade in Angola and South Africa	9
GC030	Car bombings near Madrid	6
GC031	Combats and embargo in the northern part of Iraq	59
GC032	Independence movement in Quebec	31
GC033	International sports competitions in the Ruhr area	20
GC034	Malaria in the tropics	3
GC035	Credits to the former Eastern bloc	6
GC036	Automotive industry around the Sea of Japan	0
GC037	Archeology in the Middle East	16
GC038	Solar or lunar eclipse in Southeast Asia	1
GC039	Russian troops in the southern Caucasus	16
GC040	Cities near active volcanoes	14
GC041	Shipwrecks in the Atlantic Ocean	4
GC042	Regional elections in Northern Germany	2
GC043	Scientific research in New England universities	8
GC044	Arms sales in former Yugoslavia	38
GC045	Tourism in Northeast Brazil	6
GC046	Forest fires in Northern Portugal	3
GC047	Champion league games near the Mediterranean	24
GC048	Fishing in Newfoundland and Greenland	48
GC049	ETA in France	2
GC050	Cities along the Danube and the Rhine	15

GeoTime

A geo-temporal evaluation task was introduced at NTCIR in 2008 and continued in 2009[5]. Organised by UC Berkeley, NII (and others), the aim of the evaluation was to assess the performance of systems developed to handle mixed geo-temporal information needs represented by questions, such as "When and where did George Kennan die?", "How long after the Sumatra earthquake did the tsunami hit Sri Lanka?", "When and where were the last three Winter Olympics held?", "When did direct flights between China and Taiwan begin?" and "When and where did Cyrus Vance die?". Collections were formed from English, Japanese and Korean news corpora.

[5] http://metadata.berkeley.edu/NTCIR-GeoTime/

GikiCLEF

GikiCLEF[6] was an evaluation task that ran within CLEF in 2009 (Santos *et al.*, 2010). The task was a successor to the GikiP pilot task which ran in 2008 within GeoCLEF (Santos *et al.*, 2009). The aim of the task was to evaluate systems designed to identify Wikipedia articles answering information needs requiring some form of geographical reasoning, for example "List the Italian places where Ernest Hemingway visited during his life", "Which Flemish towns hosted a restaurant with two or three Michelin stars in 2008?", "Which countries have the white, green and red colors in their national flag?" and "In which countries outside Bulgaria are there published opinions on Petar Dunov's (Beinsa Duno's) ideas?". Similar to question-answering, the information needs were in the form of questions; however, unlike question-answering the results were a list of articles containing answers rather than automatically extracted answers to the questions. Dumps of Wikipedia in the following languages were used as the document collections: Bulgarian, Dutch, English, German, Italian, Norwegian, Portuguese, Romanian and Spanish.

7.5 Evaluation in the exemplar systems

Despite the considerable effort that went into the creation of the resources described above, none of the exemplar systems we describe used these resources in their evaluation. Chief amongst the reasons for this lack of uptake of these, potentially very useful, evaluation resources are:

1. Licensing conditions associated with the corpora in question;
2. The nature of the topics developed, which often appeared to be well-suited to standard text search;
3. The emphasis on cross-language retrieval, which none of our exemplar systems engaged with; and
4. A realisation amongst many authors that the match between domain knowledge, corpora and gazetteers is central to improving performance in GIR (c.f. (Lieberman and Samet, 2012; Leveling, 2015))

Therefore, in the following we demonstrate the use of evaluation approaches for assessing the performance of two of our exemplar GIR systems, where evaluation is described in more detail. The first is the creation of a test collection (MIDR_2010) based on a Cranfield approach to assess the performance of the PIV system (Palacio *et al.*, 2010). This allows a system-oriented evaluation of different settings of the system. The second is the evaluation

[6] http://www.linguateca.pt/GikiCLEF/

of the SPIRIT system utilising both system- and user-oriented approaches to evaluation (Bucher *et al.*, 2005).

7.5.1 The PIV system

The PIV search engine is a prototype GIR system that indexes and retrieves documents based on three dimensions: topical, spatial and temporal (Palacio *et al.*, 2010). These dimensions are captured and indexed independently and the results combined to produce a single ranked list of results. The designers of the PIV system considered using existing test collections (e.g., from GeoCLEF), but none considered the three dimensions of topic, space and time.

Therefore, a custom test collection (MIDR_2010) was created that included 41 topics representing information needs reflecting the three dimensions (e.g., "Potato famine in Ireland after mid-19th century") and a collection of 5,645 paragraphs taken from eleven books published between the 18th and 20th centuries from the Aquitaine Regional Library (after scanning and optical character recognition). To form pools of documents that could be assessed for relevance, three versions of the PIV system were used to query the document collection (using the topic title) and the results aggregated to form pools. The pools for each topic were then assessed for relevance using binary judgments for each dimension (topic, space and time). A global judgment was also requested from assessors. The judgments for each dimension and overall score were summed to reflect the number of satisfied dimensions (all considered equal). This resulted in a score 0..4 from which nDCG was used to compute retrieval effectiveness (treating the score as a graded measure of relevance).

The test collection was used to assess various combinations of evidence from the different dimensions. As expected, combining results from three dimensions resulted in the highest performance (nDCG = 0.7977). The authors also carried out a topic-by-topic analysis. This is often referred to as *failure analysis* and allows closer inspection of topic performance to help identify improvements that could be made to the system.

7.5.2 The SPIRIT system

The SPIRIT project developed a GIR application that exploited both thematic and locational elements of documents available in the Web (Purves *et al.*, 2007). A prototype GIR system was developed to handle queries of the form `<theme><spatial relationship><location>`.

Both system- and user-oriented evaluation approaches were used during the project to assess the performance of individual components of the application, e.g. different relevance ranking strategies, as well as gathering user feedback on prototype interfaces. The initial approach to system-oriented evaluation involved the creation of a test collection (Bucher *et al.*, 2005). Queries included "Hotels near Horgen", "Castles east of Edinburgh" and "Castles within 50km of Cardiff". To assess relevance, a scheme was devised that assessed thematic

relevance using a binary scheme, and spatial relevance using a ternary scheme (Purves and Clough, 2006). A total of 15 topics were produced resulting in 2,228 documents judged for relevance by five assessors. To create the relevance assessments for each topic, the assessors interactively searched the SPIRIT document collection to find as many relevant documents as possible (known as interactive search and judge), judging all results returned by the system (both relevant and non-relevant). However, a number of problems emerged with this approach, the greatest being difficulty with assessing spatial relevance. This resulted in the appearance of documents whose relevance had not been previously judged and therefore inaccuracies in computing retrieval performance for various retrieval strategies.

Therefore, a different system-oriented approach was adopted in which the relevance of the top 10 documents for 38 queries and different retrieval algorithms was assessed independently by two assessors. Thematic and spatial relevance were assessed as independent dimensions on a binary scale. Thematic relevance was defined as a document that had some significant relevance to the theme and to be spatially relevant a document had to have a footprint which was considered to be similar to the query footprint, and importantly, not of a much coarser granularity. In order for a document to be considered relevant to the query, it had to be both thematically and spatially relevant. The judgments were used to compute a normalised form of Precision at 10 for each retrieval strategy (or system). The measure was normalised to account for cases where less than 10 documents were returned. Results showed that spatially-aware search outperformed text-only search for a range of queries, and especially in cases where queries went beyond simple containment.

User-oriented evaluation was used in the SPIRIT project to assess the more general usability of the system, and in particular assess different visualisations to support users' search and browse activities. Various evaluations were conducted that involved users in some manner, including more controlled task-based evaluations in a laboratory setting and an evaluation conducted remotely by volunteers. In the task-based evaluation users were asked to complete tasks which were presented as simulated work tasks (Borlund and Ingwersen, 1997). An example task (topic title and scenario) is shown in Table 7.5.

Similar to the study protocol outlined in Kelly, 2009, each user carried out a number of search tasks using the SPIRIT application. During the search, users were asked to mark documents considered as relevant. The interactions between users and the system were recorded in log files for each task. At the end of the evaluation users were asked to complete a usability questionnaire based on the Questionnaire for User Interface Satisfaction or QUIS (Chin *et al.*, 1988). Questions regarding the effectiveness of the SPIRIT system and qualitative information about user experiences were also gathered. A final post-study interview was used to gather additional comments and explore any issues raised by the users in the questionnaires. One of the problems raised during this approach to evaluation was that users often found it hard to judge the relevance of documents for topics they had not defined themselves.

Table 7.5: Example scenario used for user-oriented evaluation in the SPIRIT project

```
Topic: "Walking in London"

Scenario: "Marie and John are visiting Britain this summer.
They planned to stay only a couple of days in London, so
they want to plan their trip as much as possible before they
leave. They intend to do all their sightseeing by foot, and
want to avoid museums and other indoor attractions wherever
possible. Imagine you are Marie or John. Try and find as many
relevant documents as you can relating to walking around
London as a tourist, for example information about guided
tours, details of other peoples' experiences, etc. Bookmark
any documents that you consider to be relevant. If you feel
you have found enough relevant documents or cannot find
anymore, you can move onto the next task."
```

In a final user evaluation conducted remotely, users were asked to search using queries of their own choosing. Similar to the lab-based experiments, quantitative and qualitative data were gathered through on-line questionnaires about the SPIRIT system. In addition, users were asked to query other existing systems, such as Google Local search. This sought to gather opinions about the SPIRIT system in comparison to existing systems people were familiar with using. Overall, the most repeated remark from users concerned the speed of the system with many users commenting on the sluggishness of the system, particularly in comparison to commercial search engines. Users also commented on the importance of providing a map-based interface for GIR, particularly with regards to making it easier to judge the relevance of retrieved documents. The use of multiple evaluation methods enabled researchers to obtain a holistic view of the system throughout development.

7.6 Summary

In this chapter we have looked at ways of evaluating GIR systems, both from system- and user-oriented perspectives. Despite the efforts undertaken on evaluation so far there is still much scope for advancement, particularly in translating the methodologies used in mainstream IR to GIR. In their review of IR evaluation, Harter and Hert, 1997 also suggest a number of emerging themes in IR evaluation which are still being pursued today and should partly inform future GIR evaluation efforts: use of multiple dimensions and methods; increasing focus on evaluation from the perspectives of several IR system stakeholders; recognition that interaction must be considered in the evaluation of IR systems; the idea that IR systems and system use are embedded in layers/levels of context leading to the need to evaluate across layers and in relationship to other systems; evaluation over time; a shift to formative in addition to summative evaluation; and an increasing blurring of the boundaries

between design and evaluation activities. Recent developments in the provision of test collections for the evaluation of individual GIR components are a welcome development (Wallgrün *et al.*, 2017; Gritta *et al.*, 2017). However, despite the success of initiatives such as GeoCLEF in stimulating research, there still remains a need for usable and open resources for the more general evaluation of complete GIR systems. Collaborations with initiatives such as the placing task in MediaEval2015, which attempt to generate tasks with more specific focus on applications may suggest a potential way forward (Choi *et al.*, 2015).

8 Future Challenges

In this monograph we set out to provide an overview of research on Geographic Information Retrieval, and opened the survey by emphasising the importance of location in search. Since some of the early work described in this paper was performed, the centrality of location to many tasks related to search and retrieval of unstructured text has become increasingly obvious, as not only commercial search systems have scrambled to incorporate geographic information, but research efforts have also increasingly emphasised its importance. This shift has led to what, in computer science generally and information retrieval more specifically, is seen as a specialised domain becoming apparently mainstream. It is no longer exotic to find papers in IR displaying results on a map, or citing Tobler's classic first law of geography (Tobler, 1970).

However, despite this apparent mainstreaming the importance of geography in search is still often reduced to location, which in turn is seen as simply one of many potential features. Thus, for example, in 2012 a group of IR academics met for the Second Strategic Workshop in Information Retrieval to discuss future challenges in IR (Allan *et al.*, 2012). From our perspective one of the most striking results of the workshop was not what were listed as future research themes and challenges, but rather what is not mentioned. Thus the importance of geography in many forms of search is completely neglected other than in the form of location as an additional feature. This is perhaps all the more surprising given the repeated mention of context and mobile search across the various future challenges discussions. We would argue that time and, more particularly space, form key characteristics of users' context and also key ways of stratifying document content. Thus, though the use of location appears to have become mainstream, we believe that the treatment of geography as something different to other features has not.

This brings us to the first of the challenges that we believe is key to the development of GIR. Understanding why geography matters requires us to delve into many fields, ranging from the rather obvious geography, across spatial cognition, linguistics, computer science, geographic information science, computational geometry, information science, cartography and so on. For example, understanding spatial referents and their use is clearly key to

methods which go beyond named entity recognition to identify geographic references and analyse their relationship with text content in more sophisticated ways. Such an endeavour is inherently interdisciplinary, yet much research in GIR continues to be mono-disciplinary. Thus, we strongly argue that there is a pressing need for research in GIR to *better recognise the need for inter- and cross-disciplinary fertilisation*, especially from fields outwith the traditional domains of computer science.

Such cross-fertilisation could lead to better problem specification and thus results which are both more directly applicable and easier to evaluate. The great effort expended in the GeoCLEF evaluation effort has not, we would argue, been taken up as much as one might have hoped. One reason for this is likely the artificial specification of the problem domain, and the related lack of more general applicability of the topics specified. Better identifying tasks based on real domain needs is one obvious way of improving the impact of GIR. Potential examples include the NewsStand system, which clearly identifies a real-problem domain, and implements a system from which, presumably, user interactions data could potentially be captured (Teitler *et al.*, 2008). A further example is Brown and Baldridge's work with its direct link to the digital humanities (Brown *et al.*, 2012) or the application of the PIV system to cultural heritage in the Pyrenees (Gaio *et al.*, 2008).

One area which is the subject of much current interest is the application of machine learning approaches, and more specifically so-called deep learning to many problems in natural language processing in general. The potential application of such methods is for example discussed in the context of named entity recognition by (Melo and Martins, 2017). However, it is important here to also consider the improvement achieved by simple, well expressed rules applied to a specific domain (Leveling, 2015), or the use of a localised gazetteer (Lieberman *et al.*, 2010) and compare such approaches to machine learning-based approaches which require training data. In particular, many machine learning approaches focus on datasets where training data are readily available in the form of metadata associated with documents (e.g., Wikipedia) which may not be a realistic scenario for more general applications applied to unstructured text. Thus, we suggest that a second key challenge for future research in GIR, and more particularly georeferencing, is *reproducible publishing of methods, algorithms, datasets and results* such that approaches can be more easily compared across corpora.

More specific challenges with regard to georeferencing, some of which were proposed by Leidner and Lieberman back in 2011, but remain current include: (i) effective geoparsing and geocoding at varying levels of granularity or scale (i.e., working effectively with local-scale and global-scale data); (ii) dealing with vernacular and historical references; (iii) more intelligent ways of processing spatial language, including as spatial expressions (e.g., "40 km north of Kabul").

With respect to indexing, a key challenge lies not only in the efficient implementation of indexes (which in any case may lie more in the domain of specialists on spatial databases),

but in better specification of what is indexed. Most GIR implementations still link locations to texts in very simple ways, often representing a document either as a "bag of points" analogous to traditional "bag of words" approaches in IR, or identifying a single document scope (often as either a point or a bounding box) which then is associated with the whole document. More nuanced representations of the relationship between locations and text would in turn allow richer querying and eventually reasoning based on source materials consisting of unstructured text. One relatively simple example of such a model is the figure-ground model presented in Chapter 2 and successfully applied for example in the extraction of structured information from image captions (Hall *et al.*, 2011). We therefore argue that a further challenge is *development of theoretically well grounded methods and models for the extraction and representation of geographic information derived from unstructured text and the incorporation of such information in indexing structures.*

In the domain of ranking two key issues are apparent. Firstly, relevance in GIR is not well defined, above all when the user is considered and not only a query in from its broader, typically user-specific, context (a similar situation to that which existed in IR in general for many years). Here work on relevance of physical objects in location-based services forms a useful exception from which many ideas can potentially be transferred (Raper, 2007; De Sabbata and Reichenbacher, 2012).

Secondly, most ranking algorithms treat textual and spatial dimensions independently, and methods for their combination are often relatively simple. Although learning to rank methods certainly have potential in GIR, the lack of available data for training leads to similar issues to those discussed with respect to georeferencing. Cai, 2011 proposed a number of further challenges in relevance ranking for GIR including: (i) the need to determine how best to evaluate spatial similarity methods; (ii) the need for developing context-adaptive relevance ranking methods and understanding how context affects the boundary between relevance and non-relevant documents; (iii) the need to support human cognition and creation of spatial relevance through the provision of visual and interactive support to aid human reasoning. Janowicz *et al.*, 2011 take a step in this direction by developing a framework for specifying the semantics of similarity in GIR. We suggest that a further challenge for GIR is thus *the application of cognitively grounded methods for combination and ranking of spatial and semantically similar information.*

With regard to interfaces, GIR has taken, with some notable exceptions, a path of least resistance. Many query interfaces use simple, well-tried and trusted search boxes in the form of either purely textual input, or some combination of text and map-based input. Results are often displayed simply as points on a map, typically using as a backdrop mapping available through web-services. Little research specifically addressing GIR has focussed on the importance of the user interface, in terms of query formulation, results display or results reformulation, despite the very active associated domains of geovisualisation and cartography, which are increasingly also dealing with methods to handle unstructured

text. We consider therefore that a key challenge for GIR is to *develop cartographically and cognitively sound interfaces which better utilise Shneiderman's well tried and tested information seeking mantra (Shneiderman, 1996) in providing users with both an overview and the possibility to explore result sets through filtering and zooming.* A related challenge concerns an aspect of query formulation, in particular the automated understanding of natural language descriptions of location that include vague spatial relations, as alluded to earlier in the context of georeferencing.

The final topic addressed in this paper was evaluation. GIR has, once again, followed a similar path to IR, with most initial efforts on test-collection based evaluation focussing either on individual system components or, more rarely, complete systems. Despite the commendable efforts of the initiators of the GeoCLEF framework, there is no well-established test-collection and most GIR systems and components are evaluated in isolation. Clearly this is an unsatisfactory situation, and furthermore much evaluation of GIR has neither been task-based, nor has it evaluated user satisfaction. The latter problem is perhaps best addressed in combination with research on interfaces as discussed above. Moving to the future of evaluation both with respect to system and user-based approaches, we argue that the need for more focussed community efforts is paramount. This relates closely to reproducible publishing as discussed above, and that shared, truly open resources may well be the best way to make progress here. Beyond problems of reproducibility and sharing of resources, we emphasise the need for GIR to more carefully focus on *user-focussed evaluation, which takes into account different modes of geographic knowledge and realistic tasks.*

In closing, we would like to emphasise a thread which we hope has become apparent in this final section. We firmly believe that GIR is a truly interdisciplinary research area. By working together more effectively, we can make more progress and, perhaps more importantly, develop sound theories against which to measure our progress. The pace of development, the availability of data and the need for methods which address these issues will only increase in the coming years - thus, if we choose to take on the challenge, there is much fruitful, necessary and stimulating work to be done.

Acknowledgements

Writing this article has been a journey, and as is sometimes the case when we travel, it took a little longer than we expected as we started. We would like to thank editor Mark Sanderson for his patience and encouragement, and all three referees for their constructive and useful comments on drafts of the manuscript. They served to substantially improve the final version, although of course all errors remain ours. We would also like to thank Pia Bereuter for preparing the figures used to illustrate basic geographic concepts, and Manuel Bär for his help in preparing the final BibTex file. RSP is grateful to the Swiss National

Science Foundation Project PlaceGen (200021_149823) under whose auspices some of the reported work was carried out.

References

Adams, B., G. McKenzie, and M. Gahegan. 2015. "Frankenplace: Interactive Thematic Mapping for Ad Hoc Exploratory Search". In: *Proceedings of the 24th International Conference on World Wide Web. WWW '15*. Florence, Italy: International World Wide Web Conferences Steering Committee. 12–22. DOI: 10.1145/2736277.2741137. URL: https://doi.org/10.1145/2736277.2741137.

Ahern, S., M. Naaman, R. Nair, and J. H. Yang. 2007. "World explorer: visualizing aggregate data from unstructured text in geo-referenced collections". In: *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. ACM. 1–10.

Ahlers, D. 2013. "Assessment of the accuracy of geonames gazetteer data". In: *Proceedings of the 7th Workshop on Geographic Information Retrieval*. ACM. 74–81.

Ahmed, S. M. Z., C. McKnight, and C. Oppenheim. 2006. "A user-centred design and evaluation of IR interfaces". *Journal of Librarianship and Information Science*. 38(3): 157–172. DOI: 10.1177/0961000606063882. eprint: https://doi.org/10.1177/0961000606063882. URL: https://doi.org/10.1177/0961000606063882.

Allan, J., B. Croft, A. Moffat, and M. Sanderson. 2012. "Frontiers, Challenges, and Opportunities for Information Retrieval: Report from SWIRL 2012 the Second Strategic Workshop on Information Retrieval in Lorne". In: *ACM SIGIR Forum*. Vol. 46. No. 1. New York, NY, USA: ACM. 2–32. DOI: 10.1145/2215676.2215678. URL: http://doi.acm.org/10.1145/2215676.2215678.

Alonso, O. and S. Mizzaro. 2009. "Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment". In: *SIGIR 2009 Workshop on The Future of IR Evaluation*. 15–16.

Aloteibi, S. and M. Sanderson. 2014. "Analyzing geographic query reformulation: An exploratory study". *Journal of the Association for Information Science and Technology*. 65(1): 13–24. DOI: 10.1002/asi.22961. URL: http://dx.doi.org/10.1002/asi.22961.

Amitay, E., N. Har'El, R. Sivan, and A. Soffer. 2004. "Web-a-where: Geotagging Web Content". In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '04*. Sheffield, United Kingdom: ACM. 273–280. DOI: 10.1145/1008992.1009040. URL: http://doi.acm.org/10.1145/1008992.1009040.

Andogah, G. 2011. *Geographically Constrained Information Retrieval: Geographically Intelligent Information Retrieval*. Germany: LAP Lambert Academic Publishing.

Andogah, G., G. Bouma, and J. Nerbonne. 2012. "Every Document Has a Geographical Scope". *Data Knowl. Eng.* 81-82(Nov.): 1–20. DOI: 10.1016/j.datak.2012.07.002. URL: http://dx.doi.org/10.1016/j.datak.2012.07.002.

Andrade, L. and M. J. Silva. 2006. "Relevance Ranking for Geographic IR." In: *In Proceedings of GIR'06*.

Armitage, L. H. and P. G. B. Enser. 1997. "Analysis of user need in image archives". *Journal of Information Science*. 23(4): 287–299. DOI: 10.1177/016555159702300403. eprint: https://doi.org/10.1177/016555159702300403. URL: https://doi.org/10.1177/016555159702300403.

Axelrod, A. E. 2003. "On Building a High Performance Gazetteer Database". In: *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1. HLT-NAACL-GEOREF '03*. Stroudsburg, PA, USA: Association for Computational Linguistics. 63–68. DOI: 10.3115/1119394.1119404. URL: http://dx.doi.org/10.3115/1119394.1119404.

Baeza-Yates, R. A. and B. A. Ribeiro-Neto. 2011. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England. URL: http://www.mir2ed.org/.

Bailey, P., P. Thomas, N. Craswell, A. P. D. Vries, I. Soboroff, and E. Yilmaz. 2008. "Relevance assessment: are judges exchangeable and does it matter". In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 667–674.

Bird, S., E. Klein, and E. Loper. 2009. *Natural Language Processing with Python*. 1st. O'Reilly Media, Inc.

Borges, K. A. V., A. H. F. Laender, C. B. Medeiros, and A. S. D. Silva. 2003. "The web as a data source for spatial databases". In: *In Anais do V Brazilian Symposium on Geoinformatics, Campos do Jordão*.

Borlund, P. 2009. "User-Centred Evaluation of Information Retrieval Systems". In: *Information Retrieval: Searching in the 21st Century*. Ed. by A. Göker and D. J. John Wiley & Sons. 21–37.

Borlund, P. and P. Ingwersen. 1997. "The development of a method for the evaluation of interactive information retrieval systems". *Journal of documentation*. 53(3): 225–250.

Brisaboa, N. R., M. R. Luaces, Á. S. Places, and D. Seco. 2010. "Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based index". *GeoInformatica*. 14(3): 307–331.

Brooke, J. 1996. "SUS: a quick and dirty usability scale". In: *Usability evaluation in industry*. Ed. by P. Jordan, B. Thomas, B. Weerdmeester, and I. McClelland.

Brown, T., J. Baldridge, M. Esteva, and W. Xu. 2012. "The substantial words are in the ground and sea: computationally linking text and geography". *Texas Studies in Literature and Language*. 54(3): 324–339.

Bucher, B., P. Clough, H. Joho, R. Purves, and A. K. Syed. 2005. "Geographic IR systems: requirements and evaluation". *Proceedings of the 22nd International Cartographic Conference*. 201(2005): 11–16.

Bugayevskiy, L. and J. Snyder. 1995. *Map Projections: A Working Manual*. CRC Press.

Buscaldi, D. 2011. "Approaches to Disambiguating Toponyms". *SIGSPATIAL Special*. 3(2): 16–19. DOI: 10.1145/2047296.2047300. URL: http://doi.acm.org/10.1145/2047296.2047300.

Cai, G. 2002. "GeoVSM: An Integrated Retrieval Model for Geographic Information". In: *Geographic Information ScienceSecond International Conference, GIScience 2002*. Springer. 70–85.

Cai, G. 2011. "Relevance Ranking in Geographical Information Retrieval". *SIGSPATIAL Special*. 3(2): 33–36. DOI: 10.1145/2047296.2047304. URL: http://doi.acm.org/10.1145/2047296.2047304.

Carbonell, J. and J. Goldstein. 1998. "The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries". In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '98*. Melbourne, Australia: ACM. 335–336. DOI: 10.1145/290941.291025. URL: http://doi.acm.org/10.1145/290941.291025.

Cardoso, N. 2011. "Evaluating Geographic Information Retrieval". *SIGSPATIAL Special*. 3(2): 46–53.

Carvalho, V. R., M. Lease, and E. Yilmaz. 2011. "Crowdsourcing for search evaluation". *SIGIR Forum*. 44: 17–22.

Case, D. O. and L. M. Given. 2016. *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior. Studies in Information*. Emerald Group Publishing Limited.

Chapelle, O., D. Metlzer, Y. Zhang, and P. Grinspan. 2009. "Expected Reciprocal Rank for Graded Relevance". In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management. CIKM '09*. Hong Kong, China: ACM. 621–630. DOI: 10.1145/1645953.1646033. URL: http://doi.acm.org/10.1145/1645953.1646033.

Chen, L., G. Cong, C. S. Jensen, and D. Wu. 2013. "Spatial keyword query processing: an experimental evaluation". In: *Proceedings of the 39th international conference on Very Large Data Bases. PVLDB'13*. Trento, Italy: VLDB Endowment. 217–228. URL: http://dl.acm.org/citation.cfm?id=2448948.2448955.

Chen, Y., T. Suel, and A. Markowetz. 2006. "Efficient query processing in geographic web search engines". In: *SIGMOD Conference*. 277–288.

Chin, J. P., V. A. Diehl, and K. L. Norman. 1988. "Development of an Instrument Measuring User Satisfaction of the Human-computer Interface". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '88*. Washington, D.C., USA: ACM. 213–218. DOI: 10.1145/57167.57203. URL: http://doi.acm.org/10.1145/57167.57203.

Choi, J., C. Hauff, O. V. L. Olivier, and B. Thomee. 2015. "The placing task at mediaeval 2015". In: *MediaEval 2015, Wurzen, Germany, 14-15 September 2015; Ceur Workshop Proceedings 1436, 2015*. CEUR.

Christoforaki, M., J. He, C. Dimopoulos, A. Markowetz, and T. Suel. 2011. "Text vs. Space: Efficient Geo-search Query Processing". In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management. CIKM '11.* Glasgow, Scotland, UK: ACM. 423–432. DOI: 10.1145/2063576.2063641. URL: http://doi.acm.org/10.1145/2063576.2063641.

Cleverdon, C. W. 1991. "The significance of the Cranfield tests on index languages". In: *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '91.* 3–12.

Cleverdon, C. W., J. Mills, and M. Keen. 1966. "Factors determining the performance of indexing systems". *Aslib Cranfield Research Project Cranfield England.*

Clough, P. 2005. "Extracting Metadata for Spatially-aware Information Retrieval on the Internet". In: *Proceedings of the 2005 Workshop on Geographic Information Retrieval. GIR '05.* Bremen, Germany: ACM. 25–30. DOI: 10.1145/1096985.1096992. URL: http://doi.acm.org/10.1145/1096985.1096992.

Clough, P. D., H. Joho, and R. Purves. 2006. "Judging the Spatial Relevance of Documents for GIR". In: *Advances in Information Retrieval: 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006. Proceedings.* Ed. by M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky. Berlin, Heidelberg: Springer Berlin Heidelberg. 548–552. DOI: 10.1007/11735106_62. URL: https://doi.org/10.1007/11735106_62.

Clough, P. and M. Sanderson. 2013. "Evaluating the performance of information retrieval systems using test collections." *Information Research.* 18(2).

Cohn, A. G. and N. M. Gotts. 1996. "The 'Egg-Yolk' Representation Of Regions with Indeterminate Boundaries". In: *Proceedings, GISDATA Specialist Meeting on Geographical Objects with Undetermined Boundaries.* Francis Taylor. 171–187.

Cole, C. 2011. "A Theory of Information Need for Information Retrieval That Connects Information to Knowledge". *J. Am. Soc. Inf. Sci. Technol.* 62(7): 1216–1231. DOI: 10.1002/asi.21541. URL: http://dx.doi.org/10.1002/asi.21541.

Cong, G. and C. S. Jensen. 2016. "Querying Geo-Textual Data: Spatial Keyword Queries and Beyond". In: *Proceedings of the 2016 International Conference on Management of Data. SIGMOD '16.* San Francisco, California, USA: ACM. 2207–2212. DOI: 10.1145/2882903.2912572. URL: http://doi.acm.org/10.1145/2882903.2912572.

Cong, G., C. S. Jensen, and D. Wu. 2009. "Efficient Retrieval of the Top-k Most Relevant Spatial Web Objects". *Proc. VLDB Endow.* 2(1): 337–348. DOI: 10.14778/1687627.1687666. URL: http://dx.doi.org/10.14778/1687627.1687666.

Coventry, K. R. and S. C. Garrod. 2004. *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions. Essays in Cognitive Psychology.* Taylor & Francis.

Crandall, D., L. Backstrom, D. Huttenlocher, and J. Kleinberg. 2009. “Mapping the world’s photos”. In: *Proceedings of the 18th International Conference on World Wide Web*. ACM. 761–770.

Curran, J. R. and S. Clark. 2003. “Language Independent NER Using a Maximum Entropy Tagger”. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4. CONLL ’03*. Edmonton, Canada: Association for Computational Linguistics. 164–167. DOI: 10.3115/1119176.1119200. URL: http://dx.doi.org/10.3115/1119176.1119200.

De Felipe, I., V. Hristidis, and N. Rishe. 2008. “Keyword Search on Spatial Databases”. In: *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering. ICDE ’08*. Washington, DC, USA: IEEE Computer Society. 656–665. DOI: 10.1109/ICDE.2008.4497474. URL: http://dx.doi.org/10.1109/ICDE.2008.4497474.

De Sabbata, S., O. Alonso, and S. Mizzaro. 2012. “Classical vs. crowdsourcing surveys for eliciting geographic relevance criteria”. In: *IIR 2012 Italian Information Retrieval Workshop*. Ed. by G. Amati, C. Carpineto, and G. Semeraro. *CEUR Workshop Proceedings*. No. 835. Dipartimento di Informatica (DIB), Università di Bari "Aldo Moro". 65–72. URL: https://doi.org/10.5167/uzh-66808.

De Sabbata, S. and T. Reichenbacher. 2010. “A Probabilistic Model of Geographic Relevance”. In: *Proceedings of the 6th Workshop on Geographic Information Retrieval. GIR ’10*. Zurich, Switzerland: ACM. 23:1–23:2. DOI: 10.1145/1722080.1722109. URL: http://doi.acm.org/10.1145/1722080.1722109.

De Sabbata, S. and T. Reichenbacher. 2012. “Criteria of geographic relevance: an experimental study”. *International Journal of Geographical Information Science*. 26(8): 1495–1520.

DeLozier, G., J. Baldridge, and L. London. 2015. “Gazetteer-independent Toponym Resolution Using Geographic Word Profiles”. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI’15*. Austin, Texas: AAAI Press. 2382–2388. URL: http://dl.acm.org/citation.cfm?id=2886521.2886652.

Derungs, C. and R. S. Purves. 2014. “From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus”. *International Journal of Geographical Information Science*. 28(6): 1272–1293. DOI: 10.1080/13658816.2013.772184. eprint: http://dx.doi.org/10.1080/13658816.2013.772184. URL: http://dx.doi.org/10.1080/13658816.2013.772184.

Derungs, C. and R. S. Purves. 2016. “Mining nearness relations from an n-grams Web corpus in geographical space”. *Spatial Cognition & Computation*. 16(4): 301–322.

Ding, J., L. Gravano, and N. Shivakumar. 2000. “Computing Geographical Scopes of Web Resources”. In: *Proceedings of the 26th International Conference on Very Large Data Bases. VLDB ’00*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 545–556. URL: http://dl.acm.org/citation.cfm?id=645926.672013.

Drosou, M. and E. Pitoura. 2010. "Search Result Diversification". *SIGMOD Rec.* 39(1): 41–47. DOI: 10.1145/1860702.1860709. URL: http://doi.acm.org/10.1145/1860702.1860709.

Dunlop, M. 2000. "Reflections on Mira: Interactive evaluation in information retrieval". *Journal of the American Society for Information Science.* 51(14): 1269–1274. DOI: 10.1002/1097-4571(2000)9999:9999<::AID-ASI1042>3.0.CO;2-7.

Dykes, J., A. M. MacEachren, and M.-J. Kraak. 2005. *Exploring geovisualization.* Elsevier.

Feng, J., M. Johnston, and S. Bangalore. 2011. "Speech and multimodal interaction in mobile search". *Signal Processing Magazine, IEEE.* 28(4): 40–49.

Ferrès, D. and H. Rodriguez. 2015. "Evaluating geographical knowledge re-ranking, linguistic processing and query expansion techniques for geographical information retrieval". In: *International Symposium on String Processing and Information Retrieval.* Springer. 311–323.

Fisher, P. 2000. "Sorites paradox and vague geographies". *Fuzzy Sets and Systems.* 113(1): 7–18. DOI: https://doi.org/10.1016/S0165-0114(99)00009-3. URL: http://www.sciencedirect.com/science/article/pii/S0165011499000093.

Frontiera, P., R. Larson, and J. Radke. 2008. "A Comparison of Geometric Approaches to Assessing Spatial Similarity for GIR". *Int. J. Geogr. Inf. Sci.* 22(3): 337–360. DOI: 10.1080/13658810701626293. URL: http://dx.doi.org/10.1080/13658810701626293.

Gaio, M., C. Sallaberry, P. Etcheverry, C. Marquesuzaa, and J. Lesbegueries. 2008. "A global process to access documentsćontents from a geographical point of view". *Journal of Visual Languages & Computing.* 19(1): 3–23.

Gale, W., K. Church, and D. Yarowsky. 1992. "One Sense Per Discourse". In: *Proceedings of the Workshop on Speech and Natural Language. HLT '91.* Harriman, New York: Association for Computational Linguistics. 233–237. DOI: 10.3115/1075527.1075579. URL: http://dx.doi.org/10.3115/1075527.1075579.

Gan, Q., J. Attenberg, A. Markowetz, and T. Suel. 2008. "Analysis of Geographic Queries in a Search Engine Log". In: *Proceedings of the First International Workshop on Location and the Web. LOCWEB '08.* Beijing, China: ACM. 49–56. DOI: 10.1145/1367798.1367806. URL: http://doi.acm.org/10.1145/1367798.1367806.

Gao, S., K. Janowicz, D. R. Montello, Y. Hu, J.-A. Yang, G. McKenzie, Y. Ju, L. Gong, B. Adams, and B. Yan. 2017. "A data-synthesis-driven method for detecting and extracting vague cognitive regions". *International Journal of Geographical Information Science.* 31(6): 1245–1271.

Gey, F. C., R. R. Larson, M. Sanderson, H. Joho, P. Clough, and V. Petras. 2005. "GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview". In: *CLEF.* 908–919.

Gey, F., R. Larson, M. Sanderson, K. Bischoff, T. Mandl, C. Womser-Hacker, D. Santos, P. Rocha, and A. Montoyo. 2007. "Challenges to evaluation of multilingual geographic information retrieval in GeoCLEF". In: *The First International Workshop on Evaluating Information Access (EVIA)*. URL: http://eprints.whiterose.ac.uk/4535/.

Goodchild, M. F. 2010. "Twenty years of progress: GIScience in 2010". *Journal of Spatial Information Science*. 2010(1): 3–20.

Graham, M. and S. De Sabbata. 2015. "Mapping information wealth and poverty: the geography of gazetteers". *Environment and Planning A*. 47(6): 1254–1264.

Gritta, M., M. T. Pilehvar, N. Limsopatham, and N. Collier. 2017. "What's missing in geographical parsing?" *Language Resources and Evaluation*. Mar. DOI: 10.1007/s10579-017-9385-8. URL: https://doi.org/10.1007/s10579-017-9385-8.

Grover, C., R. Tobin, K. Byrne, M. Woollard, J. Reid, S. Dunn, and J. Ball. 2010. "Use of the Edinburgh geoparser for georeferencing digitized historical collections". *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. 368(1925): 3875–3889. DOI: 10.1098/rsta.2010.0149. eprint: http://rsta.royalsocietypublishing.org/content/368/1925/3875.full.pdf. URL: http://rsta.royalsocietypublishing.org/content/368/1925/3875.

Hall, M. M., P. D. Smart, and C. B. Jones. 2011. "Interpreting Spatial Language in Image Captions". *Cognitive Processing*. 12(1): 67–94. DOI: 10.1007/s10339-010-0385-5.

Hariharan, R., B. Hore, C. Li, and S. Mehrotra. 2007a. "Processing Spatial-Keyword (SK) Queries in Geographic Information Retrieval (GIR) Systems". In: *19th International Conference on Scientific and Statistical Database Management (SSDBM 2007)*. 16–16. DOI: 10.1109/SSDBM.2007.22.

Hariharan, R., B. Hore, C. Li, and S. Mehrotra. 2007b. "Processing Spatial-Keyword (SK) Queries in Geographic Information Retrieval (GIR) Systems". In: *Proceedings of the 19th International Conference on Scientific and Statistical Database Management. SSDBM '07*. Washington, DC, USA: IEEE Computer Society. 16–16. DOI: 10.1109/SSDBM.2007.22. URL: http://dx.doi.org/10.1109/SSDBM.2007.22.

Harman, D. 2011. *Information Retrieval Evaluation*. 1st. Morgan & Claypool Publishers.

Harter, S. P. and C. A. Hert. 1997. "Evaluation of information retrieval systems: Approaches, issues, and methods." *Annual Review of Information Science and Technology (ARIST)*. 32: 3–94.

Hearst, M. A. 2009. *Search User Interfaces*. 1st. New York, NY, USA: Cambridge University Press.

Henrich, A. and V. Luedecke. 2007. "Characteristics of Geographic Information Needs". In: *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval. GIR '07*. Lisbon, Portugal: ACM. 1–6. DOI: 10.1145/1316948.1316950. URL: http://doi.acm.org/10.1145/1316948.1316950.

Herskovits, A. 1986. *Language and Spatial Cognition: An Interdisciplinary Study of Prepositions in English*. Cambridge University Press.

Hill, L. L. 2000. "Core elements of digital gazetteers: placenames, categories, and footprints". In: *Research and advanced technology for digital libraries*. Springer. 280–290.

Hill, L. L. 2006. *Georeferencing: The Geographic Associations of Information (Digital Libraries and Electronic Publishing)*. The MIT Press.

Hill, L. L., L. Carver, M. Larsgaard, R. Dolin, T. R. Smith, J. Frew, and M.-A. Rae. 2000. "Alexandria digital library: user evaluation studies and system design". *Journal of the American Society for Information Science*. 51(3): 246–259.

Himmelstein, M. 2005. "Local Search: The Internet Is the Yellow Pages". *Computer*. 38(2): 26–34. DOI: http://doi.ieeecomputersociety.org/10.1109/MC.2005.65.

Hjaltason, G. and H. Samet. 1999. "Distance browsing in spatial databases". *ACM Transactions on Database Systems*. 24(2): 265–318.

Hobona, G., P. James, and D. Fairbairn. 2005. "An Evaluation of a Multidimensional Visual Interface for Geographic Information Retrieval". In: *Proceedings of the 2005 Workshop on Geographic Information Retrieval. GIR '05*. Bremen, Germany: ACM. 5–8. DOI: 10.1145/1096985.1096988. URL: http://doi.acm.org/10.1145/1096985.1096988.

Hoeber, O. and X. D. Yang. 2007. "User-Oriented Evaluation Methods for Interactive Web Search Interfaces". In: *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops. WI-IATW '07*. IEEE Computer Society. 239–243.

Hofmann, K., L. Li, and F. Radlinski. 2016. "Online Evaluation for Information Retrieval". *Foundations and Trends(R) in Information Retrieval*. 10(June): 1–117.

Ide, N. and J. Véronis. 1998. "Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art". *Comput. Linguist.* 24(1): 2–40. URL: http://dl.acm.org/citation.cfm?id=972719.972721.

Janowicz, K., M. Raubal, and W. Kuhn. 2011. "The semantics of similarity in geographic information retrieval". *Journal of Spatial Information Science*. 2011(2): 29–57.

Järvelin, K. 2011. "Evaluation". In: *Interactive information seeking, behaviour and retrieval*. Ed. by I. Ruthven and D. Kelly. London, UK: Facet Publishing.

Järvelin, K. and J. Kekäläinen. 2000. "IR Evaluation Methods for Retrieving Highly Relevant Documents". In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '00*. Athens, Greece: ACM. 41–48. DOI: 10.1145/345508.345545. URL: http://doi.acm.org/10.1145/345508.345545.

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: *Geographic Information Science*. Ed. by M. J. Egenhofer, C. Freksa, and H. J. Miller. Vol. 3234. *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 125–139. DOI: 10.1007/978-3-540-30231-5_9. URL: http://dx.doi.org/10.1007/978-3-540-30231-5_9.

Jones, C. B. and R. S. Purves. 2008. "Geographical Information Retrieval". *Int. J. Geogr. Inf. Sci.* 22(3): 219–228. DOI: 10.1080/13658810701626343. URL: http://dx.doi.org/10.1080/13658810701626343.

Jones, C. B., R. S. Purves, P. D. Clough, and H. Joho. 2008. "Modelling Vague Places with Knowledge from the Web". *Int. J. Geogr. Inf. Sci.* 22(10): 1045–1065. DOI: 10.1080/13658810701850547. URL: http://dx.doi.org/10.1080/13658810701850547.

Karimzadeh, M., W. Huang, S. Banerjee, J. O. Wallgrün, F. Hardisty, S. Pezanowski, P. Mitra, and A. M. MacEachren. 2013. "GeoTxt: A Web API to Leverage Place References in Text". In: *Proceedings of the 7th Workshop on Geographic Information Retrieval. GIR '13*. Orlando, Florida: ACM. 72–73. DOI: 10.1145/2533888.2533942. URL: http://doi.acm.org/10.1145/2533888.2533942.

Karney, C. F. F. 2013. "Algorithm for Geodesics". *Journey of Geodetics*. 87(1): 43–55.

Kelly, D. 2009. "Methods for evaluating interactive information retrieval systems with users". *Foundations and Trends in Information Retrieval*. 3(1-2): 1–224.

Keßler, C., K. Janowicz, and M. Bishr. 2009. "An agenda for the next generation gazetteer: Geographic information contribution and retrieval". In: *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in Geographic Information Systems*. ACM. 91–100.

Khodaei, A., C. Shahabi, and C. Li. 2010. "Hybrid Indexing and Seamless Ranking of Spatial and Textual Features of Web Documents". In: *DEXA*. 450–466.

Khodaei, A., C. Shahabi, and C. Li. 2012. "SKIF-P: a point-based indexing and ranking of web documents for spatial-keyword search". *GeoInformatica*. 16(3): 563–596.

Kinney, K. A., S. B. Huffman, and J. Zhai. 2008. "How evaluator domain expertise affects search result relevance judgments". In: *Proceedings of the 17th ACM conference on Information and knowledge management. CIKM '08*. ACM. 591–598.

Kinsella, S., V. Murdock, and N. O'Hare. 2011. ""I'M Eating a Sandwich in Glasgow": Modeling Locations with Tweets". In: *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents. SMUC '11*. Glasgow, Scotland, UK: ACM. 61–68. DOI: 10.1145/2065023.2065039. URL: http://doi.acm.org/10.1145/2065023.2065039.

Kreveld, M., I. Reinbacher, A. Arampatzis, and R. Zwol. 2005. "Developments in Spatial Data Handling: 11th International Symposium on Spatial Data Handling". In: Berlin, Heidelberg: Springer Berlin Heidelberg. Chap. Distributed Ranking Methods for Geographic Information Retrieval. 231–243. DOI: 10.1007/3-540-26772-7_18. URL: http://dx.doi.org/10.1007/3-540-26772-7_18.

Laere, O. V., S. Schockaert, V. Tanasescu, B. Dhoedt, and C. B. Jones. 2014. "Georeferencing Wikipedia Documents Using Data from Social Media Sources". *ACM Trans. Inf. Syst.* 32(3): 12:1–12:32. DOI: 10.1145/2629685. URL: http://doi.acm.org/10.1145/2629685.

Landau, B. and R. Jackendoff. 1993. ""What" and "where" in spatial language and spatial cognition". *Behavioral and Brain Sciences*. 16(2): 217–238.

Larson, R. R. 1996. "Geographic Information Retrieval and Spatial Browsing". *GIS and Libraries: Patrons, Maps and Spatial Information*. Apr.: 81–124. Ed. by L. Smith and M. Gluck.

Larson, R. R. 2011. "Ranking Approaches for GIR". *SIGSPATIAL Special*. 3(2): 37–41. DOI: 10.1145/2047296.2047305. URL: http://doi.acm.org/10.1145/2047296.2047305.

Larson, R. R. and P. Frontiera. 2004. "Spatial Ranking Methods for Geographic Information Retrieval (GIR) in Digital Libraries". In: *Research and Advanced Technology for Digital Libraries, 8th European Conference, ECDL 2004, Bath, UK, September 12-17, 2004, Proceedings*. 45–56.

Leidner, J. L. 2006. "An evaluation dataset for the toponym resolution task". *Computers, Environment and Urban Systems*. 30(4): 400–417.

Leidner, J. L. 2008. *Toponym Resolution in Text : Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Boca Raton, FL, USA: Universal Press.

Leidner, J. L. and M. D. Lieberman. 2011. "Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language". *SIGSPATIAL Special*. 3(2): 5–11. DOI: 10.1145/2047296.2047298. URL: http://doi.acm.org/10.1145/2047296.2047298.

Leveling, J. 2015. "Tagging of Temporal Expressions and Geological Features in Scientific Articles". In: *Proceedings of the 9th Workshop on Geographic Information Retrieval. GIR '15*. Paris, France: ACM. 6:1–6:10. DOI: 10.1145/2837689.2837701. URL: http://doi.acm.org/10.1145/2837689.2837701.

Levinson, S. C. 2003a. "Frames of reference". In: *Space in Language and Cognition: Explorations in Cognitive Diversity. Language Culture and Cognition*. Cambridge University Press. 24–61. DOI: 10.1017/CBO9780511613609.003.

Levinson, S. C. 2003b. *Space in language and cognition: Explorations in cognitive diversity*. Cambridge: CUP.

Lewis, J. R. 1995. "IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use". *International Journal of Human-Computer Interaction*. 7(1): 57–78.

Li, H. 2011. "A Short Introduction to Learning to Rank". *IEICE Transactions on Information and Systems*. E94.D(10): 1854–1862. DOI: 10.1587/transinf.E94.D.1854.

Li, H., R. K. Srihari, C. Niu, and W. Li. 2003. "InfoXtract Location Normalization: A Hybrid Approach to Geographic References in Information Extraction". In: *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1. HLT-NAACL-GEOREF '03*. Stroudsburg, PA, USA: Association for Computational Linguistics. 39–44. DOI: 10.3115/1119394.1119400. URL: http://dx.doi.org/10.3115/1119394.1119400.

Li, Z., K. C. K. Lee, B. Zheng, W.-C. Lee, D. L. Lee, and X. Wang. 2011. "IR-Tree: An Efficient Index for Geographic Document Search." *IEEE Transactions on Knowledge and Data Engineering*. 23(4): 585–599. URL: http://dblp.uni-trier.de/db/journals/tkde/tkde23.html#LiLZLLW11.

Lieberman, M. D., H. Samet, and J. Sankaranarayanan. 2010. "Geotagging with local lexicons to build indexes for textually-specified spatial data". In: *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*. 201–212. DOI: 10.1109/ICDE.2010.5447903.

Lieberman, M. D., H. Samet, J. Sankaranarayanan, and J. Sperling. 2007. "STEWARD: Architecture of a Spatio-textual Search Engine". In: *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems. GIS '07*. Seattle, Washington: ACM. 25:1–25:8. DOI: 10.1145/1341012.1341045. URL: http://doi.acm.org/10.1145/1341012.1341045.

Lieberman, M. and H. Samet. 2012. "Adaptive Context Features for Toponym Resolution in Streaming News". In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '12*. Portland, Oregon, USA: ACM. 731–740. ISBN: 978-1-4503-1472-5. DOI: 10.1145/2348283.2348381. URL: http://doi.acm.org/10.1145/2348283.2348381.

Liu, T.-Y. 2009. "Learning to Rank for Information Retrieval". *Foundations and Trends in Information Retrieval*. 3(3): 225–331. DOI: 10.1561/1500000016. URL: http://dx.doi.org/10.1561/1500000016.

Longley, P. A., M. F. Goodchild, D. J. Maguire, and D. W. Rhind. 2015. *Geographic information science and systems*. John Wiley & Sons.

Lowe, D. G. 2004. "Distinctive image features from scale-invariant keypoints". *International journal of computer vision*. 60(2): 91–110.

MacEachren, A. M. 1995. *How maps work: representation, visualization, and design*. Guilford Press.

Mackaness, W. A., A. Ruas, and L. T. Sarjakoski. 2011. *Generalisation of geographic information: cartographic modelling and applications*. Elsevier.

Mandl, T. 2011. "Evaluating GIR: Geography-oriented or User-oriented?" *SIGSPATIAL Special*. 3(2): 42–45.

Mandl, T., P. Carvalho, G. M. Di Nunzio, F. Gey, R. R. Larson, D. Santos, and C. Womser-Hacker. 2008a. "GeoCLEF 2008: the CLEF 2008 cross-language geographic information retrieval track overview". In: *Evaluating Systems for Multilingual and Multimodal Information Access*. Springer. 808–821.

Mandl, T., F. Gey, G. (Di Nunzio), N. Ferro, M. Sanderson, D. Santos, and C. Womser-Hacker. 2008b. "An Evaluation Resource for Geographic Information Retrieval". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Ed. by N. C. (Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, and D. Tapias. http://www.lrec-conf.org/proceedings/lrec2008/. Marrakech, Morocco: European Language Resources Association (ELRA).

Manguinhas, H., B. Martins, and J. Borbinha. 2008. "A geo-temporal web gazetteer integrating data from multiple sources". In: *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*. IEEE. 146–153.

Manning, C. D., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Mao, J., Y. Liu, K. Zhou, J.-Y. Nie, J. Song, M. Zhang, S. Ma, J. Sun, and H. Luo. 2016. "When Does Relevance Mean Usefulness and User Satisfaction in Web Search?" In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '16*. Pisa, Italy: ACM. 463–472. DOI: 10.1145/2911451.2911507. URL: http://doi.acm.org/10.1145/2911451.2911507.

Marchionini, G. 2006. "Exploratory Search: From Finding to Understanding". *Commun. ACM*. 49(4): 41–46. DOI: 10.1145/1121949.1121979. URL: http://doi.acm.org/10.1145/1121949.1121979.

Mark, D. M., D. Comas, M. . Egenhofer, S. M. Freundschuh, M. D. Gould, and J. Nunes. 1995. "Evaluating and refining computational models of spatial relations through cross-linguistic human-subjects testing". In: *Spatial Information Theory A Theoretical Basis for GIS*. Vol. 988/1995. *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg. 553–568.

Martins, B. 2011. "A supervised machine learning approach for duplicate detection over gazetteer records". In: *GeoSpatial Semantics*. Springer. 34–51.

Martins, B. and P. Calado. 2010. "Learning to Rank for Geographic Information Retrieval". In: *Proceedings of the 6th Workshop on Geographic Information Retrieval. GIR '10*. Zurich, Switzerland: ACM. 21:1–21:8. DOI: 10.1145/1722080.1722107. URL: http://doi.acm.org/10.1145/1722080.1722107.

Martins, B., M. J. Silva, and M. S. Chaves. 2005. "Challenges and Resources for Evaluating Geographical IR". In: *Proceedings of the 2005 Workshop on Geographic Information Retrieval. GIR '05*. Bremen, Germany: ACM. 65–69.

Al-Maskari, A., M. Sanderson, and P. Clough. 2007. "The Relationship Between IR Effectiveness Measures and User Satisfaction". In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '07*. Amsterdam, The Netherlands: ACM. 773–774. ISBN: 978-1-59593-597-7. DOI: 10.1145/1277741.1277902. URL: http://doi.acm.org/10.1145/1277741.1277902.

McCurley, K. S. 2001. “Geospatial Mapping and Navigation of the Web”. In: *Proceedings of the 10th International Conference on World Wide Web. WWW ’01*. Hong Kong, Hong Kong: ACM. 221–229. DOI: 10.1145/371920.372056.

Melo, F. and B. Martins. 2017. “Automated Geocoding of Textual Documents: A Survey of Current Approaches”. *Transactions in GIS*. 21(1): 3–38.

Mikheev, A., M. Moens, and C. Grover. 1999. “Named Entity Recognition Without Gazetteers”. In: *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics. EACL ’99*. Bergen, Norway: Association for Computational Linguistics. 1–8. DOI: 10.3115/977035.977037. URL: http://dx.doi.org/10.3115/977035.977037.

Montello, D. R. 2016. “Cognition and Spatial Behavior”. In: *International Encyclopedia of Geography: People, the Earth, Environment and Technology*. John Wiley & Sons, Ltd. DOI: 10.1002/9781118786352.wbieg0498. URL: http://dx.doi.org/10.1002/9781118786352.wbieg0498.

Montello, D. R., M. F. Goodchild, J. Gottsegen, and P. Fohl. 2003. “Where’s Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries”. *Spatial Cognition & Computation*. 3(3): 185–104.

Morimoto, Y., M. Aono, M. E. Houle, and K. S. McCurley. 2003. “Extracting spatial knowledge from the web”. In: *2003 Symposium on Applications and the Internet, 2003. Proceedings*. 326–333. DOI: 10.1109/SAINT.2003.1183066.

Morville, P. and L. Rosenfeld. 2006. *Information Architecture for the World Wide Web*. O’Reilly Media, Inc.

Nadeau, D. and S. Sekine. 2007. “A Survey of Named Entity Recognition and Classification”. *Journal of Linguisticae Investigationes*. 30(1): 1–20. URL: http://nlp.cs.nyu.edu/sekine/papers/li07.pdf.

Nielsen, J. 1993. *Usability Engineering*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

O’Hare, N. and V. Murdock. 2013. “Modeling locations with social media”. *Information Retrieval*. 16(1): 30–62.

O’Sullivan, D. and D. Unwin. 2014. *Geographic information analysis*. John Wiley & Sons.

Opach, T., I. Golebiowska, and S. I. Fabrikant. 2013. “How Do People View Multi-Component Animated Maps?” *The Cartograhic Journal*. online(Oct.). DOI: dx.doi.org/10.1179/1743277413Y.0000000049.

Overell, S. and S. Rüger. 2008. “Using co-occurrence models for placename disambiguation”. *International Journal of Geographical Information Science*. 22(3): 265–287.

Palacio, D., G. Cabanac, C. Sallaberry, and G. Hubert. 2010. “On the Evaluation of Geographic Information Retrieval Systems: Evaluation Framework and Case Study”. *Int. J. Digit. Libr.* 11(2): 91–109. DOI: 10.1007/s00799-011-0070-z. URL: http://dx.doi.org/10.1007/s00799-011-0070-z.

Palacio, D., C. Derungs, and R. Purves. 2015. "Development and evaluation of a geographic information retrieval system using fine grained toponyms". *Journal of Spatial Information Science*. (11): 1–29.

Pasley, R. C., P. D. Clough, and M. Sanderson. 2007. "Geo-tagging for Imprecise Regions of Different Sizes". In: *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval. GIR '07*. Lisbon, Portugal: ACM. 77–82. DOI: 10.1145/1316948.1316969. URL: http://doi.acm.org/10.1145/1316948.1316969.

Purves, R. S. and P. D. Clough. 2006. "Judging spatial relevance and document location for Geographic Information Retrieval". In: *In Proceedings of 4th International Conference on Geographic Information Science (GIScience 2006)*. 159–164.

Purves, R. S., P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, and B. Yang. 2007. "The Design and Implementation of SPIRIT: A Spatially Aware Search Engine for Information Retrieval on the Internet". *Int. J. Geogr. Inf. Sci.* 21(7): 717–745. DOI: 10.1080/13658810601169840. URL: http://dx.doi.org/10.1080/13658810601169840.

Purves, R. S., A. Edwardes, and M. Sanderson. 2008. "Describing the where–improving image annotation and search through geography". In: *Proceedings of the workshop on Metadata Mining for Image Understanding (MMIU 2008)*. Sheffield.

Raper, J. 2007. "Geographic relevance". *Journal of Documentation*. 63(6): 836–852.

Rapp, R. H. 1993. "Geometric geodesy, part II, Technical report," *tech. rep.* Ohio State Univ. URL: http://hdl.handle.net/1811/24409.

Rauch, E., M. Bukatin, and K. Baker. 2003. "A Confidence-based Framework for Disambiguating Geographic Terms". In: *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1. HLT-NAACL-GEOREF '03*. Stroudsburg, PA, USA: Association for Computational Linguistics. 50–54. DOI: 10.3115/1119394.1119402. URL: http://dx.doi.org/10.3115/1119394.1119402.

Recchia, G. and M. M. Louwerse. 2013. "A Comparison of String Similarity Measures for Toponym Matching." In: *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place. COMP '13*. Orlando FL, USA: ACM. 54:54–54:61. DOI: 10.1145/2534848.2534850. URL: http://doi.acm.org/10.1145/2534848.2534850.

Reichenbacher, T., S. De Sabbata, R. S. Purves, and S. I. Fabrikant. 2016. "Assessing geographic relevance for mobile search: A computational model and its validation via crowdsourcing". *Journal of the Association for Information Science and Technology*. 67(11): 2620–2634. DOI: 10.1002/asi.23625. URL: http://dx.doi.org/10.1002/asi.23625.

Robertson, S. E. 1981. "The methodology of information retrieval experiment". In: *Information retrieval experiment*. Butterworths. 9–31.

Robertson, S. E. and M. M. Hancock-Beaulieu. 1992. "On the Evaluation of IR Systems". *Inf. Process. Manage.* 28(4): 457–466. DOI: 10.1016/0306-4573(92)90004-J. URL: http://dx.doi.org/10.1016/0306-4573(92)90004-J.

Robertson, S. and H. Zaragoza. 2009. "The Probabilistic Relevance Framework: BM25 and Beyond". *Foundations and Trends Information Retrieval.* 3(4): 333–389. DOI: 10.1561/1500000019. URL: http://dx.doi.org/10.1561/1500000019.

Rocha-Junior, J. B., O. Gkorgkas, S. Jonassen, and K. Nørvåg. 2011. "Efficient Processing of Top-k Spatial Keyword Queries". In: *Proceedings of the 12th International Conference on Advances in Spatial and Temporal Databases. SSTD'11.* Minneapolis, MN: Springer-Verlag. 205–222. URL: http://dl.acm.org/citation.cfm?id=2035253.2035270.

Rodden, K., H. Hutchinson, and X. Fu. 2010. "Measuring the User Experience on a Large Scale: User-centered Metrics for Web Applications". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '10.* Atlanta, Georgia, USA: ACM. 2395–2398. DOI: 10.1145/1753326.1753687. URL: http://doi.acm.org/10.1145/1753326.1753687.

Roller, S., M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. 2012. "Supervised text-based geolocation using language models on an adaptive grid". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* Association for Computational Linguistics. 1500–1510.

Russell-Rose, T. and T. Tate. 2013. *Designing the Search Experience: The Information Architecture of Discovery.* 1st. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Sanderson, M. 2010. "Test Collection Based Evaluation of Information Retrieval Systems". *Foundations and Trends in Information Retrieval.* 4(4): 247–375. DOI: 10.1561/1500000009. URL: http://dx.doi.org/10.1561/1500000009.

Sanderson, M. and J. Kohler. 2004. "Analyzing geographic queries". In: *Proceedings of the Workshop on Geographic Information Retrieval.* Sheffield.

Santos, D., L. M. Cabral, C. Forascu, P. Forner, F. C. Gey, K. Lamm, T. Mandl, P. Osenova, A. Peñas, Á. Rodrigo, J. M. Schulz, Y. Skalban, and E. F. T. K. Sang. 2010. "GikiCLEF: Crosscultural Issues in Multilingual Information Access." In: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).* Ed. by N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias. European Languages Resources Association (ELRA). 2346–2353.

Santos, D., N. Cardoso, P. Carvalho, I. Dornescu, S. Hartrumpf, J. Leveling, and Y. Skalban. 2009. "GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia". In: *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum*. Ed. by C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. F. Jones, M. Kurimo, T. Mandl, A. Peñas, and V. Petras. Vol. 5706. *Lecture Notes in Computer Science (LNCS)*. Springer. 894–905.

Santos, R. L. T., C. Macdonald, and I. Ounis. 2015. "Search Result Diversification". *Foundations and Trends in Information Retrieval*. 9(1): 1–90. DOI: 10.1561/1500000040. URL: http://dx.doi.org/10.1561/1500000040.

Saracevic, T. 1995. "Evaluation of Evaluation in Information Retrieval". In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '95*. Seattle, Washington, USA: ACM. 138–146. DOI: 10.1145/215206.215351.

Saracevic, T. 1996. "Relevance reconsidered". In: *Information science: Integration in perspectives. Proceedings of the Second Conference on Conceptions of Library and Information Science*. 201–218.

Schockaert, S. 2011. "Vague Regions in Geographic Information Retrieval". *SIGSPATIAL Special*. 3(2): 24–28. DOI: 10.1145/2047296.2047302. URL: http://doi.acm.org/10.1145/2047296.2047302.

Sehgal, V., L. Getoor, and P. D. Viechnicki. 2006. "Entity resolution in geospatial data integration". In: *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*. ACM. 83–90.

Shatford, S. 1986. "Analyzing the subject of a picture: a theoretical approach". *Cataloging & classification quarterly*. 6(3): 39–62.

Shaw, B., J. Shea, S. Sinha, and A. Hogue. 2013. "Learning to Rank for Spatiotemporal Search". In: *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. WSDM '13*. Rome, Italy: ACM. 717–726. DOI: 10.1145/2433396.2433485. URL: http://doi.acm.org/10.1145/2433396.2433485.

Shneiderman, B. 1996. "The eyes have it: A task by data type taxonomy for information visualizations". In: *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE. 336–343.

Shneiderman, B., D. Byrd, and W. B. Croft. 1998. "Sorting out Searching: A User-interface Framework for Text Searches". *Commun. ACM*. 41(4): 95–98. DOI: 10.1145/273035.273069. URL: http://doi.acm.org/10.1145/273035.273069.

Smart, P. D., C. B. Jones, and F. A. Twaroch. 2010. "Multi-source Toponym Data Integration and Mediation for a Meta-Gazetteer Service". In: *Geographic Information Science: 6th International Conference, GIScience 2010, Zurich, Switzerland, September 14-17, 2010. Proceedings*. Ed. by S. I. Fabrikant, T. Reichenbacher, M. van Kreveld, and C. Schlieder. Berlin, Heidelberg: Springer Berlin Heidelberg. 234–248. DOI: 10.1007/978-3-642-15300-6_17. URL: https://doi.org/10.1007/978-3-642-15300-6_17.

Smith, D. A. and G. Crane. 2001. "Disambiguating geographic names in a historical digital library". In: *Research and Advanced Technology for Digital Libraries*. Springer. 127–136.

Smith, D. A. and G. S. Mann. 2003. "Bootstrapping Toponym Classifiers". In: *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1. HLT-NAACL-GEOREF '03*. Stroudsburg, PA, USA: Association for Computational Linguistics. 45–49. DOI: 10.3115/1119394.1119401. URL: http://dx.doi.org/10.3115/1119394.1119401.

Speriosu, M. and J. Baldridge. 2013. "Text-Driven Toponym Resolution using Indirect Supervision". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*. 1466–1476. URL: http://aclweb.org/anthology/P/P13/P13-1144.pdf.

Spink, A., B. J. Jansen, D. Wolfram, and T. Saracevic. 2002. "From e-sex to e-commerce: Web search changes". *Computer*. 35(3): 107–109. DOI: 10.1109/2.989940.

Spink, A., D. Wolfram, M. B. J. Jansen, and T. Saracevic. 2001. "Searching the web: The public and their queries". *Journal of the Association for Information Science and Technology*. 52(3): 226–234.

Stokes, N., Y. Li, A. Moffat, and J. Rong. 2008. "An empirical study of the effects of NLP components on Geographic IR performance". *Int. J. Geogr. Inf. Sci.* 22(3): 247–264.

Su, L. T. 2003. "A comprehensive and systematic model of user evaluation of Web search engines: I. Theory and background". *J. Am. Soc. Inf. Sci. Technol.* 54(13): 1175–1192. DOI: 10.1002/asi.10303. URL: http://dx.doi.org/10.1002/asi.10303.

Sutcliffe, A. and M. Ennis. 1998. "Towards a cognitive theory of information retrieval". *Interacting with Computers*. 10(3): 321–351. {HCI} and Information Retrieval. DOI: http://dx.doi.org/10.1016/S0953-5438(98)00013-7. URL: http://www.sciencedirect.com/science/article/pii/S0953543898000137.

Talmy, L. 1983. "How Language Structures Space". In: *Spatial Orientation*. New York: Plenum. 225–282.

Tang, J. and M. Sanderson. 2010. "Evaluation and User Preference Study on Spatial Diversity". In: *Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010.Proceedings*. Ed. by C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger, and K. Van Rijsbergen. Berlin, Heidelberg: Springer Berlin Heidelberg. 179–190. DOI: 10.1007/978-3-642-12275-0_18. URL: https://doi.org/10.1007/978-3-642-12275-0_18.

Teitler, B. E., M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. 2008. "NewsStand: A New View on News". In: *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. GIS '08*. Irvine, California: ACM. 18:1–18:10. DOI: 10.1145/1463434.1463458. URL: http://doi.acm.org/10.1145/1463434.1463458.

Thomas, P. and D. Hawking. 2006. "Evaluation by Comparing Result Sets in Context". In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management. CIKM '06*. Arlington, Virginia, USA: ACM. 94–101. DOI: 10.1145/1183614.1183632. URL: http://doi.acm.org/10.1145/1183614.1183632.

Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region". *Economic Geography*. 46: 234–240. URL: http://www.jstor.org/stable/143141.

Uryupina, O. 2003. "Semi-supervised Learning of Geographical Gazetteers from the Internet". In: *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1. HLT-NAACL-GEOREF '03*. Stroudsburg, PA, USA: Association for Computational Linguistics. 18–25. DOI: 10.3115/1119394.1119397. URL: http://dx.doi.org/10.3115/1119394.1119397.

Vaid, S., C. B. Jones, H. Joho, and M. Sanderson. 2005. "Spatio-textual Indexing for Geographical Search on the Web". In: *Advances in Spatial and Temporal Databases: 9th International Symposium, SSTD 2005, Angra dos Reis, Brazil, August 22-24, 2005. Proceedings*. Ed. by C. Bauzer Medeiros, M. J. Egenhofer, and E. Bertino. Berlin, Heidelberg: Springer Berlin Heidelberg. 218–235. DOI: 10.1007/11535331_13. URL: https://doi.org/10.1007/11535331_13.

Vakkari, P. 2012. "Evaluating Interactive Information Retrieval Systems". *Revista PRISMA.COM*. 2012(19): 1–15. URL: http://revistas.ua.pt/index.php/prismacom/article/view/2410.

Vakkari, P. and S. Huuskonen. 2012. "Search effort degrades search output but improves task outcome". *Journal of the American Society for Information Science and Technology*. 63(4): 657–670. DOI: 10.1002/asi.21683. URL: http://dx.doi.org/10.1002/asi.21683.

Van Rijsbergen, C. J. 1979. *Information Retrieval*. 2nd. Newton, MA, USA: Butterworth-Heinemann.

Vaughan, M. W. and M. L. Resnick. 2006. "Search User Interfaces: Best Practices and Future Visions". *J. Am. Soc. Inf. Sci. Technol.* 57(6): 777–780. DOI: 10.1002/asi.v57:6. URL: http://dx.doi.org/10.1002/asi.v57:6.

Voorhees, E. M. and D. K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press.

Wallgrün, J. O., M. Karimzadeh, A. M. MacEachren, and S. Pezanowski. 2017. "GeoCorpora: building a corpus to test and train microblog geoparsers". *International Journal of Geographical Information Science*. 0(0): 1–29. DOI: 10.1080/13658816.2017.1368523. eprint: http://dx.doi.org/10.1080/13658816.2017.1368523. URL: http://dx.doi.org/10.1080/13658816.2017.1368523.

Wang, C., X. Xie, L. Wang, Y. Lu, and W.-Y. Ma. 2005. "Web Resource Geographic Location Classification and Detection". In: *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web. WWW '05*. Chiba, Japan: ACM. 1138–1139. DOI: 10.1145/1062745.1062907. URL: http://doi.acm.org/10.1145/1062745.1062907.

Wang, W. and K. Stewart. 2015. "Creating spatiotemporal semantic maps from web text documents". In: *Space-Time Integration in Geography and GIScience*. Springer. 157–174.

White, R. W. 2016. *Interactions with Search Systems*. Cambridge University Press. DOI: 10.1017/CBO9781139525305. URL: http://dx.doi.org/10.1017/CBO9781139525305.

Wilkening, J. and S. I. Fabrikant. 2013. "How Users Interact With a 3D Geo-Browser under Time Pressure". *Cartography and Geographic Information Science*. 40: 40–52.

Wilson, M. L. 2011. *Search User Interface Design*. Morgan & Claypool Publishers.

Wing, B. P. and J. Baldridge. 2011. "Simple supervised document geolocation with geodesic grids". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics. 955–964.

Woodruff, A. G. and C. Plaunt. 1994. "GIPSY: Automated Geographic Indexing of Text Documents". *J. Am. Soc. Inf. Sci.* 45(9): 645–655. DOI: 10.1002/(SICI)1097-4571(199410)45:9<645::AID-ASI2>3.0.CO;2-8.

Wu, D., G. Cong, and C. S. Jensen. 2012. "A Framework for Efficient Spatial Web Object Retrieval". *The VLDB Journal*. 21(6): 797–822. DOI: 10.1007/s00778-012-0271-0. URL: http://dx.doi.org/10.1007/s00778-012-0271-0.

Xiao, X., Q. Luo, Z. Li, X. Xie, and W.-Y. Ma. 2010. "A Large-scale Study on Map Search Logs". *ACM Trans. Web*. 4(3): 8:1–8:33. DOI: 10.1145/1806916.1806917. URL: http://doi.acm.org/10.1145/1806916.1806917.

Yan, H., S. Ding, and T. Suel. 2009. "Inverted Index Compression and Query Processing with Optimized Document Ordering". In: *Proceedings of the 18th International Conference on World Wide Web. WWW '09*. Madrid, Spain: ACM. 401–410. DOI: 10.1145/1526709.1526764. URL: http://doi.acm.org/10.1145/1526709.1526764.

Zaila, Y. L. and D. Montesi. 2015. "Geographic Information Extraction, Disambiguation and Ranking Techniques". In: *Proceedings of the 9th Workshop on Geographic Information Retrieval. GIR '15*. Paris, France: ACM. 11:1–11:7. DOI: 10.1145/2837689.2837695. URL: http://doi.acm.org/10.1145/2837689.2837695.

Zandbergen, P. A. 2008. "A comparison of address point, parcel and street geocoding techniques". *Computers, Environment and Urban Systems*. 32(3): 214–232.

Zhai, C. X., W. W. Cohen, and J. Lafferty. 2003. "Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval". In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '03*. Toronto, Canada: ACM. 10–17. DOI: 10.1145/860435.860440. URL: http://doi.acm.org/10.1145/860435.860440.

Zhang, C., Y. Zhang, W. Zhang, and X. Lin. 2013. “Inverted linear quadtree: Efficient top k spatial keyword search”. In: *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. 901–912. DOI: 10.1109/ICDE.2013.6544884.

Zhou, Y., X. Xie, C. Wang, Y. Gong, and W.-Y. Ma. 2005. “Hybrid Index Structures for Location-based Web Search”. In: *Proceedings of the 14th ACM International Conference on Information and Knowledge Management. CIKM ’05*. Bremen, Germany: ACM. 155–162. DOI: 10.1145/1099554.1099584. URL: http://doi.acm.org/10.1145/1099554.1099584.