Chapter 1
Describing and Selecting Collections of Georeferenced Media Items in Peer-to-Peer Information Retrieval Systems

Daniel Blank
University of Bamberg, Germany

Andreas Henrich
University of Bamberg, Germany

ABSTRACT

The ever-increasing amount of media items on the World Wide Web and on private devices leads to a strong need for adequate indexing and search techniques. Trends such as personal media archives, social networks, mobile devices with huge storage space, and networks with high bandwidth capacities make distributed solutions and in particular Peer-to-Peer (P2P) Information Retrieval (IR) systems attractive. On the other hand, when designing effective media search applications, various search criteria have to be addressed. Hereby, geospatial information is frequently used as well as other criteria, such as text, audio or visual media content, and date and time information.

In this chapter, the authors outline how collections of georeferenced media items can be indexed and searched in P2P IR systems. They discuss different types of P2P IR systems and focus in detail on an approach based on collection description and selection techniques. This approach tries to adequately describe and select collections of georeferenced media items. Finally, the authors discuss its broad applicability in various application fields.

DOI: 10.4018/978-1-4666-0945-7.ch001

INTRODUCTION

In recent years, the availability—and with it the usefulness—of geospatial metadata has increased dramatically. Digital cameras and mobile phones are nowadays often equipped with GPS sensors at affordable cost. Hence, such devices are able to capture georeferenced information in the personal lives of millions of people from all over the world. In addition, geo-tagging tools with rich user interfaces have emerged in different domains and large geo-tagging initiatives try to georeference textual resources such as in case of Wikipedia. As a consequence, an increased importance of geospatial information in the context of search can be recognized.

Obviously, *geospatial information* is not the only search criterion. When searching for media items other criteria such as *textual content*, *timestamps*, and *(low-level) audio or visual content information* can be used as well—often in an integrative way. A combination of these criteria can allow for the effective retrieval of text, image, audio, and video documents.

As we all know, the amount of media items on the World Wide Web and on private devices steadily increases. Service providers such as Flickr, YouTube, or Facebook (Beaver, et al., 2010) have to maintain huge hardware infrastructures in order to keep up with the tremendous increase in data volumes. So far, it is unclear if existing server-centered solutions will also suit our needs in the future. Hence, a need for alternative indexing and search techniques might arise.

Peer-to-Peer (P2P) Information Retrieval (IR) systems consist of computers from all over the world. These computers can act as both clients and servers. By applying a scalable P2P IR protocol, a "service of equals" for the administration of media items can be established in contrast to existing client/server-based solutions. No expensive infrastructure has to be maintained and idle computing power in times of inactivity can be used to maintain, analyze, and enrich media items. P2P IR systems offer the benefit that media items can remain on individual devices since there is no need for storing them on remote servers hosted by third party service providers. Crawling which consumes large amounts of web traffic (Bockting & Hiemstra, 2009) can thus be avoided. In addition, dependency from service providers acting as informational gatekeepers can be reduced, because they are no longer able to decide which information can be retrieved or accessed and which cannot. In times of a strong market concentration in internet search and social network applications as well as public debates addressing the privacy of data, P2P IR could offer some benefits.

As our primary use case, georeferenced images are administered in a P2P IR system. The images of a certain user are stored locally on the user's personal device(s) and a scalable P2P IR protocol is applied in order to facilitate retrieval. An image can hereby be described by various criteria: textual metadata, (low-level) visual content features, a timestamp, and a geographic coordinate. Personal media collections containing multiple images can thus be represented by corresponding collection descriptions allowing for efficient and effective collection selection when processing a given query. We assume in the following that at least some of the images of a peer are geo-tagged. A resource description capturing the geographic footprint of an image collection can thus be generated from the set of georeferenced images. In this chapter, we focus on geospatial query processing. In particular, we address geospatial k nearest neighbor (k-NN) queries—finding the k closest media items according to a given query location.

In literature, many approaches for P2P IR can be found. The following section entitled *Peer-to-Peer Information Retrieval Systems for Geospatial Search* will give an overview on different types of P2P IR systems and outline how georeferenced media items can be indexed in a P2P setting in general. We additionally describe how a comprehensive indexing of the abovementioned search criteria (geospatial, textual, date and time, and audio or

visual content information) can be achieved. In addition, we discuss associated consequences for query processing in a distributed scenario.

A P2P IR system which might use the above-mentioned criteria and which is based on resource (i.e. collection) selection is presented in detail in section *Resource Selection based on Descriptions of Geospatial Footprints*. Resource selection techniques are applied to determine a ranking of promising resources based on descriptions of their geospatial footprint. Peers are contacted in ranked order to retrieve the media items with the closest locations according to a user-given query location. It is important to note that the core of our P2P IR system—its resource description and selection techniques addressing georeferenced media items—is not limited to P2P IR systems. Different application fields and usage scenarios of these resource description and selection techniques are therefore discussed at the end of this chapter.

In a *Future Work* section, we will focus on different open aspects w.r.t. geospatial resource description and selection before we end with a *Conclusion* section.

PEER-TO-PEER INFORMATION RETRIEVAL SYSTEMS FOR GEOSPATIAL SEARCH

P2P IR systems are mostly classified as being *structured* or *unstructured* overlay networks. In addition, we introduce the distinction of *data-independent* and *data-dependent* overlays as a secondary classification criterion to reflect if, for example, a peer's content or query profiles affect overlay generation. This distinction provides helpful insights in order to pinpoint different characteristics in a more organized way. In the following subsections, we will briefly discuss unstructured and structured as well as hybrid and super-peer approaches.

Unstructured Topologies

Data-independent: Main protocols in this group are PlanetP (Cuenca-Acuna, et al., 2003) and its extension Rumorama (Müller, et al., 2005). In Rumorama, a peer sees the network as a single, small PlanetP network (called subnet) with connections to other peers, which see other PlanetP subnets. Each peer can choose the size of its subnet according to local processing power and bandwidth capacity. Within a subnet, a peer knows data summaries of all other peers in the subnet. Gossiping techniques are used to disseminate the data summaries. In a subnet, summary-based resource selection allows for semantic query routing. Additionally, a peer maintains a small set of links pointing to neighboring peers in other subnets in order to be able to forward queries outside the boundaries of its own subnet.

In its original form, peers are assigned to subnets arbitrarily, i.e. independent of the peers' content. However, Rumorama can be easily extended by a grouping of peers—similar to the content-dependent overlays described in the following—for example by content or by geospatial proximity of the host, which operates the peer. A further benefit of the resource descriptions is their possibility of being visualized. Thus, they might be applied for interactive retrieval, e.g. by providing—with low bandwidth requirements—a visual overview of peer data for a large number of peers.

Routing indexes in various forms represent aggregated information in an unstructured P2P IR system maintained at a peer for all its neighboring peers in order to decide in which direction queries should be forwarded. Initially designed for one-dimensional values in order to avoid network flooding in the early days of P2P computing, they have for example been extended to allow for multi-dimensional queries. Here, bounding boxes are used to summarize the content of neighboring peers (for references cf. Doulkeridis, et al., 2009).

Data-dependent: Many Semantic Overlay Networks (SONs; for references and a detailed description cf. Doulkeridis, et al., 2010) are data-dependent, unstructured P2P networks. Here, the content of a peer's data or information about past queries defines a peer's place in the network. Thus, summaries of a peer's content or query profiles are needed. Two types of links are usually maintained in such systems: short links grouping peers with similar content or query profiles into so called "Clusters of Interest" (COIs) and long links which are established between different COIs. During query processing, the query has to be forwarded to the most promising COI(s). Clustering, classification, and gossiping techniques can be applied in order to form COIs.

Indexing of Multiple Criteria in Unstructured Topologies: Many of the unstructured, data-independent P2P IR systems are based on resource descriptions which summarize a single peer's content, summarize the content of multiple peers reachable when following a certain direction, or for example summarize information about past queries. In general, it is possible to apply one summary type and a corresponding resource selection technique per feature type. Feature-specific peer rankings can be combined by applying an algorithm for the merging of ranked lists (Belkin, et al., 1995; Ilyas, et al., 2008). As an alternative for creating independent resource descriptions per search criterion, summaries and resource selection algorithms integrating multiple feature types are also possible (cf. Hariharan, et al., 2008).

In section *Resource Selection based on Descriptions of Geospatial Footprints* we will focus on resource description and selection techniques for geospatial information. The design of resource description and selection techniques for textual data is for example addressed in Cuenca-Acuna (2003). Summaries for high-dimensional feature vectors in order to summarize content-based media features are outlined in Blank et al. (2007).

In data-dependent unstructured networks multiple search criteria can be addressed when forming the COIs. Since many approaches are based on a similarity measure between resources' content, different criteria can be integrated when determining the similarity of peers in order to group them together. Alternatively, multiple overlays might be maintained, i.e. one overlay per search criterion.

Structured Topologies

Data-independent: Structured P2P IR systems are based on distributed indexing structures. Distributed Hash Tables (DHTs) represent the most prominent class member. Every peer in the network is usually responsible for a certain range of the feature space. Thus, when entering the network or updating local content, index data has to be transferred to remote peers according to the peers' responsibilities. In case of data-independent, structured P2P IR systems, terms (cf. Bender, et al., 2005) or high-dimensional feature vectors for content-based image retrieval (cf. Novak, et al., 2008; Lupu, et al., 2007; Vu, et al., 2009) are usually mapped to one-dimensional or multi-dimensional keys. They can be indexed in a classical DHT such as Chord (Stoica, et al., 2001) or CAN (Ratnasamy, et al., 2001) respectively. It has to be noted that there is a large variety of such P2P protocols. Very detailed information with references pointing to relevant research articles can be found in Shen et al. (2010).

Data-dependent: SONs—as described above—can also be implemented on top of a DHT in order to enhance query routing (Doulkeridis, et al., 2010). Various forms of clustering, classification together with gossiping techniques can be applied in order to establish links to peers with similar content.

Indexing of Multiple Criteria in Structured Topologies: In structured, data-independent systems, correlations between different criteria are difficult to exploit when indexing multiple feature types (e.g. geospatial and image content information). If we for example assume an image from the Sahara Desert with shades of beige sand and

blue sky, different peers might be responsible for indexing the geospatial and the image content information. Thus, when distributing the index data of the Sahara image, querying for it, or removing it from the network, (at least) two different peers have to be contacted. Within SONs, the simultaneous indexing of multiple criteria would again require the definition of a similarity between peers' content by combining for example geospatial and image content information. Alternatively, multiple overlays might be maintained.

Hybrid and Super-Peer Approaches

In unstructured P2P IR systems, a peer only administers index data of media items, which belong to its user. Thus, when entering the network or updating media items, full index data does not have to be transferred to remote peers. Peer autonomy is better respected compared to structured networks (Doulkeridis, et al., 2010). On the other hand, structured systems offer query processing with logarithmic cost. In order to reduce the load imposed on the network when inserting new media items in structured systems, super-peer architectures (Papapetrou, et al., 2007) as well as DHT-based indexing of compact data summaries instead of full index data have been proposed (cf. Lupu, et al., 2007).

In general, there is a convergence of structured and unstructured P2P IR systems with many hybrid approaches. We have for example evaluated an approach where index data is transferred amongst peers in certain rounds in order to make peers more focused and—as a consequence—summaries more selective. More selective summaries with peers having specialized on a certain range of the feature space lead to more efficient resource selection (Eisenhardt, et al., 2008).

There is plenty of work addressing super-peer architectures (for references cf. Doulkeridis, et al., 2009). They are designed in order to overcome some limitations of "true" P2P IR systems and make use of increased capabilities such as storage capacity, processing power, or available network bandwidth. Often, concepts known from "true" P2P IR systems are extended and transferred to super-peer architectures. Also within super-peer networks the convergence of different approaches can be seen. Doulkeridis et al. (2009) for example apply multi-dimensional routing indexes on a super-peer level and additionally group similar super-peers close together in order to allow for improved query routing.

RESOURCE SELECTION BASED ON DESCRIPTIONS OF GEOSPATIAL FOOTPRINTS

Geospatial Resource Description and Selection in Rumorama-Like P2P IR Systems

It is important to note that the resource selection techniques presented in the following are not restricted to data-independent, unstructured P2P IR systems such as Rumorama. The summaries can also be applied in data-dependent, unstructured P2P IR systems when forming COIs and in structured networks in order to be indexed for example in a DHT. In addition, summaries can be used by super-peers for selecting "normal" peers or other super-peers. Here, it might be possible to relax the strict compliance of some design parameters. A system based on super-peers might for example allow for less space efficient resource descriptions in order to achieve better retrieval performance.

In this section, we will present four different resource description and corresponding selection techniques for geospatial information. Further application fields of the analyzed geospatial resource description and selection techniques outside the P2P IR context are also possible. We will focus on this aspect in the next section which is entitled *Resource Descriptions and Selection Techniques in different Application Fields*.

In this chapter, we outline and extend earlier work from Blank and Henrich (2009, 2010) mainly according to two major directions. First, we analyze peer ranking selectivity in more detail by not only providing the average fraction of contacted peers as the main performance measure. We analyze the distribution of the fraction of contacted peers over the set of queries, which gives additional insights. Second, we show that also on a global scale (i.e. based on a data collection of geo-tagged Flickr images from different areas of the world), it is sufficient in our scenario to approximate distances by the use of Euclidean distance when searching for the 20 nearest neighbors according to a given query location.

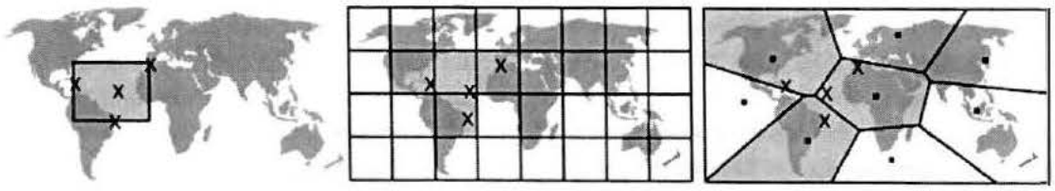
We use the terms *image* and *document* interchangeably in order to refer to concrete media items. This is due to the fact that our experiments in Blank and Henrich (2009, 2010) and also in this very chapter are based on image documents. Nevertheless, we believe that our approach can be extended to other media types. Of course, by restricting our analysis on images, we implicitly allow only a single lat/lon-coordinate per media item. This might be different in case of textual documents for example where several locations could be referenced within a single document. A detailed analysis of these and related challenges could be part of future work.

Bounding Boxes (BB)

When using bounding boxes as resource descriptions, every peer computes a minimum bounding box over the geospatial coordinates of its image collection (refer to. Figure 1). We encode a latitude/longitude-pair (for short: lat/lon-pair) with 8 bytes, 4 for latitude and 4 for longitude. Therefore, we require 16 bytes of raw data for the bounding box (i.e. two lat/lon-pairs, e.g. the lower left and upper right corner).

Peer ranking is performed as follows. If a peer p_a contains the query location in its bounding box whereas peer p_b does not, peer p_a is ranked higher than peer p_b and vice versa. In case the query location lies in the bounding box of both peers p_a and p_b, the size of a peer (i.e. the number of documents a peer administers) is used as an additional criterion. Peers with more documents are ranked higher. In earlier studies, several alternatives have been evaluated (cf. Blank & Henrich, 2010) and this strategy turned out to be beneficial. If neither peer p_a nor peer p_b contain the query within its bounding box, the peer with the smaller minimum distance from the query location to its bounding box is preferred. We assume a spherical model of the earth with a radius of 6371 km. If not stated otherwise, we use Haversine distance (Sinnott, 1984) to compute the distance between two points on the sphere.

Figure 1. Visualizing summary creation for BB (left), $GRID_4$ (middle), and HFS_8/UFS_8 (right: ■ corresponds to reference points). Four documents indicated as x are geo-tagged.



Of course, it is also possible to represent the geospatial footprint of a peer by multiple bounding boxes. Becker et al. (1991) present an algorithm for summarizing a set of bounding boxes by two bounding boxes which minimize the area that is covered. Chen et al. (2006) propose several threshold-based algorithms to split a single bounding box into several smaller ones in order to reduce the space within a bounding box which is not covered by any index data. These approaches demonstrate that there is optimization potential w.r.t. the representation of geo-regions in geospatial indexing. However, the mentioned approaches stick with bounding boxes and the approaches presented later on in this chapter might be interesting alternatives in this application field. A detailed comparison might be part of future work.

Grid-Based Summaries ($GRID_r$)

In a second approach, the lat/lon-coordinates are mapped to a grid (see. Figure 1). A parameter r represents the number of rows of the grid. The number of columns is twice the number of rows since longitude range is twice as big as latitude range. The range of a grid cell (in degrees) is determined by $180° / r = 360° / (2r)$ in the latitude and longitude domain respectively. This simplified view is for example also applied in Dolin et al. (1997) and results in non-uniform grid cell sizes on the sphere. We gain selectivity and retrieval performance by increasing the number of grid cells at the price of additional storage overhead partially compensated through compression techniques (cf. section *Experiments*). Every grid cell is represented by a single bit. If one or more image locations fall into a certain cell, the corresponding bit is set to 1. Otherwise, it remains 0. Bit positions in the summary are determined horizontally from left to right and from bottom to top. Effects of alternative strategies on compression could be evaluated in future work.

During peer ranking the grid cell containing the query location is determined. If peer p_a has an image in this cell whereas peer p_b has not, peer p_a is ranked higher than peer p_b and vice versa. We also consider neighboring grid cells. If either both or none of peer p_a and peer p_b have an image located within the cell containing the query location, the GRID approach considers the neighboring cells recursively until a ranking decision can be made. So, in the first round the ranking decision is always based on a single cell; in the second round it is in most cases based on 1+8=9 cells and in the third round on 1+8+16=25 cells and so on. This is not always the case since there might be no neighboring cells in a certain direction, e.g. as soon as a cell in the north or south is reached. Of course, at the 180° meridian we assume that there is no boundary and neighborhood relations are valid in both directions. The ranking criterion in every round is the number of grid cells containing one or more image location(s)—the more the better.

Of course, there are alternative approaches in literature for mapping the spherical coordinates to a grid for example in order to achieve uniform grid cell sizes on the sphere (Putman, et al., 2007). In various domains, cubed grids (for references cf. Putman, et al., 2007), or triangular meshes (Szalay, et al., 2005) have been proposed. These techniques might also be considered in future work, although we do not expect large improvements compared to our naïve approach—neither in terms of ranking selectivity nor in terms of reduced summary sizes. We gain ranking selectivity by increasing the number of grid cells. On the other hand, this increase does only lead to a sublinear increase of resource description sizes since compression techniques are applied.

Highly Fine-Grained Summaries (HFS_n)

This approach is based on resource descriptions originally designed for summarizing visual content information of images, for example the color distribution or texture of an image (Blank, et al., 2007). A set of n predetermined image locations are used as reference points. This set of reference

points is known to all peers and built from external sources (a Gazetteer is used in our case). How to adequately obtain the reference points is outlined in the remainder of this chapter. So far, it is sufficient to note that every image location of a peer's local image collection is assigned to the closest reference point according to Haversine distance (cf. Figure 1, right). Hereby, a cluster histogram is computed counting how many image locations of a peer's collection are closest to a certain reference point, i.e. cluster center c_j ($1 \leq j \leq n$).

Peer ranking is performed as follows. Reference points c_j are sorted in ascending order according to Haversine distance to the query. The first element of the sorted list L corresponds to the cluster center being closest to the query. Peers with more documents in this so called query cluster are ranked higher than peers with fewer documents in the query cluster. If two peers administer the same amount of documents in the analyzed cluster, the next element out of L is chosen and both peers are ranked according to the number of documents within the very cluster. This procedure continues until either a ranking decision can be made, which favors one peer over the other, or the end of L is reached. In the latter case, a random decision would be made.

Ultra Fine-Grained Summaries (UFS_n)

In contrast to HFS, UFS are based on a bit vector with the bit at position j indicating if center j is the closest center to one or more of a peer's image locations. Therefore, we obtain a bit vector of size n. Of course, there is some loss of information when switching from HFS to UFS with n staying constant. However, UFS have the potential of resulting in more space efficient resource descriptions. Potentially, this allows for more reference points being used, which might result in similar or even improved retrieval performance compared to HFS. Among other aspects, this is evaluated in Blank and Henrich (2010) and will be briefly summarized in the following experimental section. Before doing so, we will describe the data collections we use.

The abovementioned resource description techniques as well as variations and combinations of them (e.g. combining grid and bounding box based representations) are also used in the context of multidimensional and metric index structures (Samet, 2006). The decision of choosing the best subtree is similar to the resource selection problem. Summaries in the P2P context correspond for example to aggregations maintained in the nodes of a tree, for example bounding boxes in the case of an R-tree (Guttman, 1984). We therefore point the interested reader to Samet (2006).

Experiments

Data Collections

Two data collections of geo-tagged images are used in our experiments:

1. **Geoflickr:** During the year 2007 a large amount of publicly available images was crawled which had been uploaded to Flickr (http://www.flickr.com). In our scenario, every Flickr user operates a peer of its own. We therefore assign images to peers by means of the Flickr user ID. All of the crawled images are geo-tagged. After some data cleansing the Geoflickr collection consists of 406450 geo-tagged images from 5951 different users/peers.
2. **Geograph:** Geograph (http://www.geograph.org.uk/) "aims to collect geographically representative photographs and information for every square kilometre of Great Britain and Ireland." We downloaded the geo-tagged images and distributed them to peers again in a user-centric approach. In our scenario every Geograph participant operates a single peer: 2609 peers administer 246937 images and thus image locations in total.

The distribution of the number of images per peer is displayed for both collections in Figure 2. For both collections the distribution of the number of images per peer is very skew which is typical for many P2P settings (Cuenca-Acuna, et al., 2003). Few peers administer large amounts of the collection. On the other hand, there are many peers which store only few images. A more detailed analysis of the distribution of the peer sizes, i.e. the number of images per peer, can be found in Blank and Henrich (2010).

A visualization of the geographic distribution of the image locations can be found in Blank and Henrich (2010), too. The Geoflickr collection consists of photos taken in various parts of the world with hotspots in North America, Europe, and Japan. In contrast, images of the Geograph collection are limited to the UK and Ireland with images more densely located around urban areas such as London.

Experimental Settings

In our experiments, we use 200 image locations as queries. These are randomly selected from the underlying data collection. The query locations are visualized in Figure 3. For HFS and UFS where the outcome of the experiments is affected by the selection of reference points, we run at least ten experiments with the 200 queries each. Since we do not remove the image with the query location, it is—on average—more likely that a big peer contributes to the retrieval result than a small peer because—on average—it is more likely to choose the query from a big peer than from a small peer. An additional strategy for selecting the queries is analyzed in Blank and Henrich (2010).

Space efficiency of different resource descriptions is measured by analyzing average summary sizes. For compressing the summaries we apply Java's gzip implementation with default parameter values. Our measurements include serializa-

Figure 2. Number of images per peer for the Geoflickr and Geograph collection

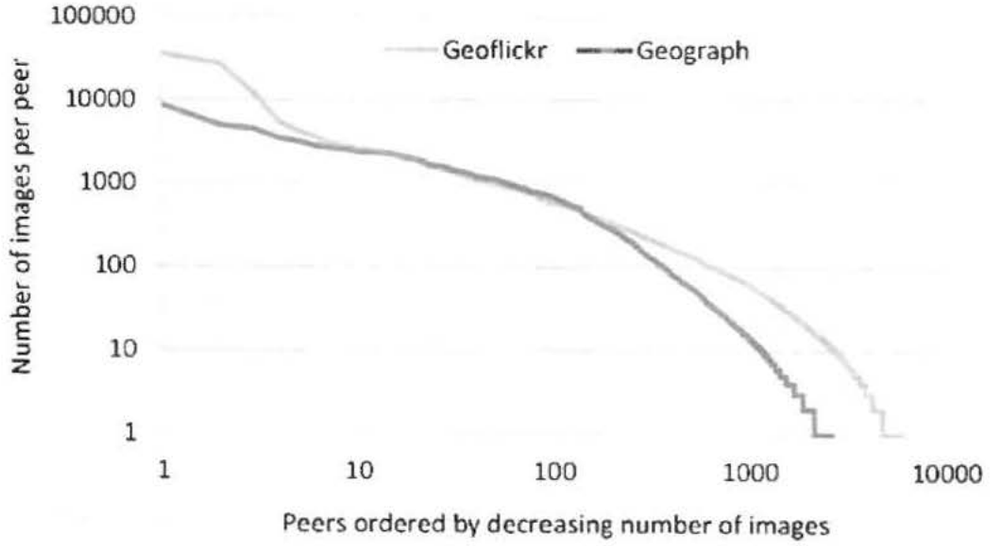
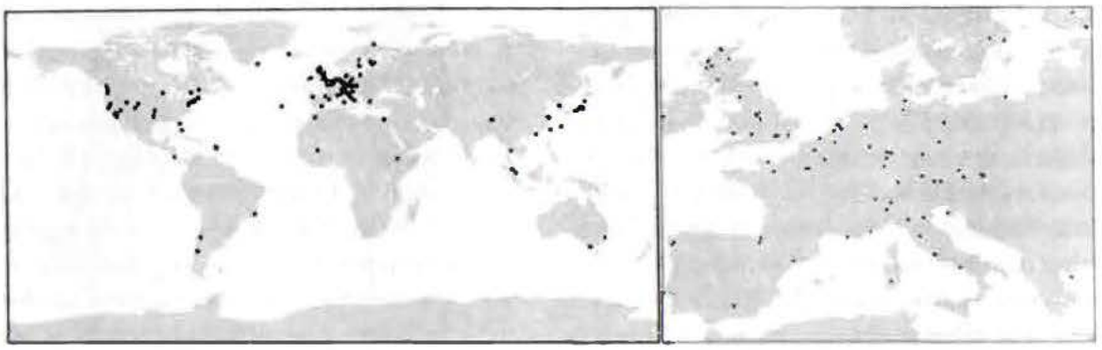


Figure 3. Image locations used as queries for Geoflickr



tion overhead necessary in order to distribute the resource descriptions in the network. The product of average summary size times the number of peers indicates the network load, which is imposed by a single round of gossiping (i.e. every peer sends its summary once to all other peers in the PlanetP-like subnet).

In earlier studies (Blank & Henrich, 2010), in order to measure peer ranking selectivity we determine the fraction of peers which needs to be contacted on average to retrieve a certain fraction of the top-k image locations ($k = 20$). In this chapter, we take a closer look at ranking selectivity by analyzing the distribution of the fraction of contacted peers over all queries.

The top-k geo-locations are computed using Vincenty distance (Vincenty, 1975). Since we are interested in the selectivity of the resource description and selection techniques, we analyze all of a peer's image locations as soon as it is contacted, because the top-k image locations of a peer determined using Haversine distance might differ from the top-k image locations computed using Vincenty distance. In a real-world application, only the top-k image locations will be transferred (together with some additional information such as peer ID, etc.)

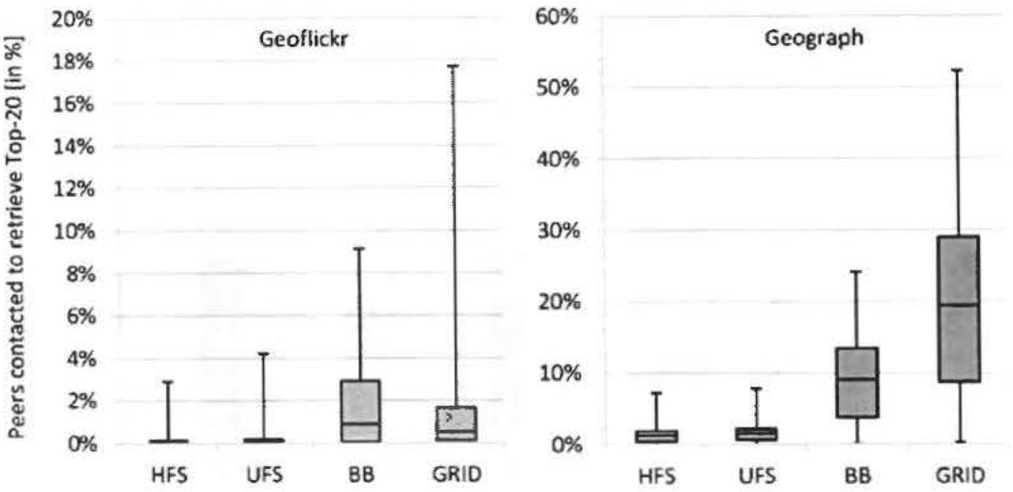
Empirical Analysis of Approaches

Peer ranking selectivity for both collections is displayed in Figure 4 with the help of boxplots including minimum and maximum values. More experimental results with a focus on the average number of peers, which are contacted during query processing, can be found in Blank and Henrich (2010). For reasons of brevity, we focus on a setting with $n = 8192$ for UFS and HFS here. For the GRID approach we partition the data space in 32 rows and 64 columns. Thus, in an uncompressed way, GRID will also result in a bit vector of size 8192.

From Figure 4 can be observed that HFS performs slightly better than UFS. This seems reasonable since HFS encodes frequency information which can be beneficial for peer ranking. If we assume that the query is closest to a certain reference point c^*, then it is obvious to contact peer p_a before peer p_b if peer p_a has assigned more image locations to c^* than peer p_b. In case of UFS, this information gets lost. Here, peer p_b with fewer image locations assigned to c^* might be selected before peer p_a. Although there might be cases where such a strategy will lead to better ranking selectivity, these cases will be exceptional.

For the Geoflickr collection (Figure 4, left), both HFS and UFS perform better than BB and GRID. Interquartile ranges of UFS and HFS are

Figure 4. Ranking selectivity for different approaches



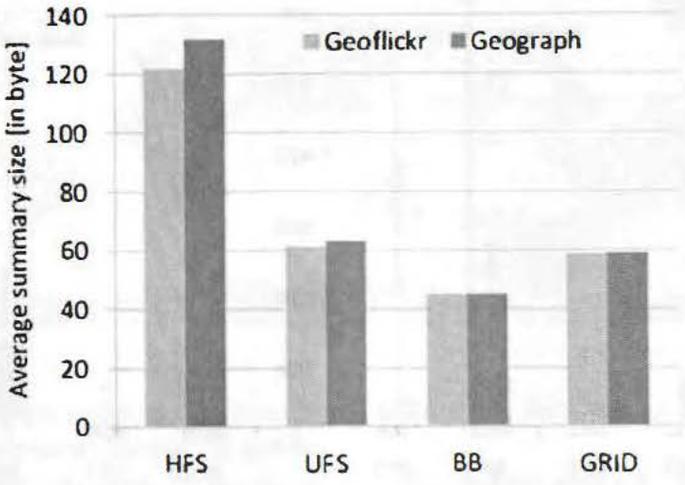
much smaller than for GRID and especially BB. GRID offers better ranking selectivity than BB. In addition to the average number of contacted peers (cf. Blank & Henrich, 2010), also median and 75th percentile of GRID are clearly below corresponding values for BB. Also interquartile range is smaller. Nevertheless, at least for some queries, it is very difficult for GRID to offer adequate ranking selectivity as indicated by the maximum value of almost 18%. This is the case for queries, which lie in a very populated grid cell, i.e. a grid cell where many peers assign documents to. We therefore looked at queries, which offer poor ranking selectivity for GRID in more detail. It can be observed that many of these queries lie in cells where many of the documents reside, i.e. cells, which contain metropolises such as London for example.

In case of Geograph (Figure 4, right), the grid is not adapted to the boundaries of the United Kingdom (of course this can easily be done). We did not adapt it in order to show the effects of a skew distribution of geospatial image locations on a global scale. HFS and UFS are better suited for such scenarios than GRID, because they better adapt to the data which is used. Also for HFS and UFS reference points are chosen on a global scale and are not restricted to the boundaries of the UK. They are selected from a Gazetteer and correspond to various locations from all over the world.

Summaries based on a single bounding box per peer lack peer ranking selectivity and seem to be too coarse in order to be able to compete with HFS or UFS, but, bounding boxes can be represented in a very space efficient way. They require only 45 byte per resource description as can be observed from Figure 5.

For both collections, HFS affords—on average—approximately twice as big resource descriptions as UFS. If we take a closer look at the Geoflickr collection, average summary sizes of UFS are slightly bigger than in case of GRID. On the other hand, further experiments show that even GRID with more cells (for example 96 rows and 192 columns) cannot outperform UFS in terms of ranking selectivity. $GRID_{96}$ contacts on average more peers than UFS_{8192} in order to retrieve the Top-20 with a bigger average summary size.

Figure 5. Average summary sizes for different approaches



A more detailed analysis of ranking selectivity for HFS and UFS is displayed in Figure 6. It visualizes the number of contacted peers for different values of n. Medians in case of HFS are always smaller than in case of UFS. It can be confirmed that HFS outperforms UFS. But, performance gaps decrease with increasing values of n. Differences in peer ranking selectivity diminish with increasing n, since the corresponding histograms become more and more similar with many zeros and some summary bin values set to 1. Of course, for HFS, the values of some summary bins might still be bigger than 1, but with increasing n this becomes rarer and rarer. Also interquartile ranges for HFS and UFS become more and more similar when increasing n. This indicates that ranking selectivity of HFS and UFS equal more and more.

In the following, we will limit ourselves to the analysis of UFS since this approach seems to offer a good compromise between ranking selectivity and summary size. In order to further trade-off these two factors under the influence of other aspects, we refer to the general cost model presented in Blank and Henrich (2010) for a more detailed analysis.

Obtaining the Reference Points

For obtaining the reference points, we employ Geonames gazetteer (http://www.geonames.org/) as well as United Nations' per country statistics obtained through Worldmapper (http://www.worldmapper.org/). Within this chapter, we focus on statistics about men's income, Gross Domestic Product (GDP), population and WWW usage. Various other statistics are evaluated in Blank and Henrich (2010). Based on the statistics we proportionally select the number of reference points from a certain country. Reference points are selected amongst all populated places of a certain country at random. So, for example, if x% of the world's GDP comes from a certain country, x% of the reference points are randomly chosen amongst all populated places of the specific country.

Figure 7 shows ranking selectivity in terms of the average fraction of peers which are contacted in order to retrieve the top-20 image locations. We plot UFS results for different values of n. An analysis of HFS offers similar characteristics and is omitted here for reasons of brevity. We can see that the strategy based on GDP performs best. This confirms the finding in Blank and Henrich

Figure 6. Ranking selectivity of HFS and UFS for different values of n

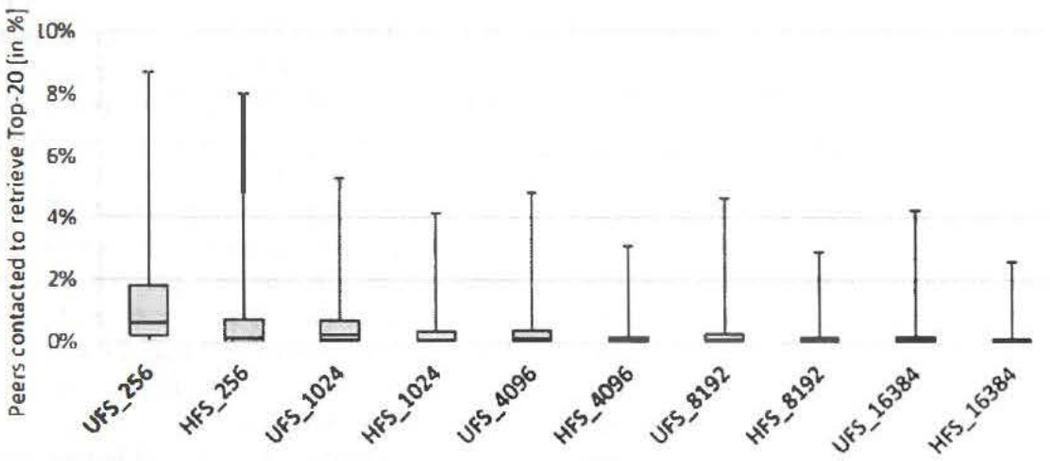
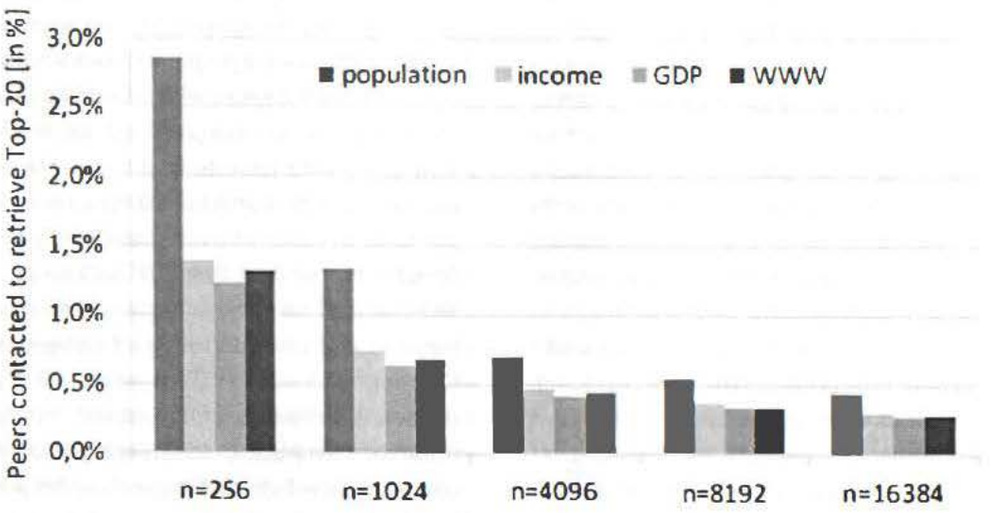


Figure 7. Choosing reference points for UFS from external sources



(2010) where only small values of n were analyzed. It should be noted here that we used the GDP-based approach in the preceding experiments of this chapter.

Analysis of Distance Measures

In our experiments, the 20 closest image locations according to a given query location are always computed using Vincenty distance (Vincenty, 1975) in order to achieve good accuracy, since we are interested in the "true" nearest neighbors. On the other hand, in our algorithms we have used Haversine distance (Sinnott, 1984) which is computationally more efficient than Vincenty distance at the price of computing less accurate distance values. Whereas Vincenty distance is based on an ellipsoidal shape of the earth, Haversine distance assumes a spherical model. In addition to Haversine and Vincenty distance, we analyzed two versions of Euclidean distance (the use of lat/lon-pairs as 2-dimensional vectors as well as 3-dimensional vectors resulting from a projection of the lat/lon-coordinates onto a sphere). These distances might be used in order to further approximate distance values and speed-up query processing. As for Haversine distance, we use a rather coarse earth radius of approximately 6731 km. We observed in experiments that all approaches offer similar retrieval performance. Hence, it is sufficient to apply Euclidean distance.

In the following section, we will describe how geospatial resource description and selection techniques, which were presented and evaluated in the preceding sections, could be applied in different application fields. By doing so we will show that their use is not limited to the P2P IR domain.

Resource Descriptions and Selection Techniques in Different Application Fields

In addition to P2P IR, geospatial resource summarization and selection techniques can also be used in traditional distributed IR applications. Personal meta-search is a novel application of distributed IR, where all the online resources of a person are queried. Web users frequently administer various e-mail accounts, bookmarks of web pages, image collections, databases, etc. These resources are typically heterogeneous in size, media type and update frequency (Thomas & Hawking, 2009) possibly requiring space efficient and at the same time selective (geospatial) resource descriptions. Thus, applications might be built providing a unified search service (similar to a meta-search engine) over all these resources. This could prevent web users from the time consuming task of querying all resources "manually."

Spatial Data Infrastructures (SDIs; for a general description see for example Nogueras-Iso et al. [2005]) might also be an interesting application field for geospatial resource description and selection techniques. Chen et al. (2010) identify three critical problems with current solutions which are for example based on the Catalogue Service for the Web (CSW; http://www.opengeospatial.org/standards/cat). First, server-based solutions offer the problem of a single point of failure. Second, it is the task of the users who search for services to identify the proper portal. Third, users who want to provide data have to find a suitable portal to register to, which might also be a time consuming task. In order to overcome these issues, it is broadly recognized that P2P technology might be beneficial for the discovery of geospatial web services and the publishing of geographic data (Xiujun, et al., 2006; Chen, et al., 2010). What remains is the question for the best P2P technology in the context of SDIs. Due to the high maintenance costs of structured solutions with nodes joining or leaving the network, Chen et al. (2010) argue for a hybrid solution. Initially nodes join an unstructured network and only if the node stays in the network for a longer time period, it can move to the structured network. While the administration of lat/lon-coordinates in a structured network might for example be based on a quadtree-like (Chen, et al., 2010) or R^+-tree-like (Xiujin, et al., 2006) structure, or a space-filling curve based mapping

to one dimensional values (Memon, et al., 2009), the unstructured part of the network might use a resource description and selection based approach for web service discovery. Here, rather than minimum bounding boxes such as in case of (Xiujin, et al., 2006), grid-based or Voronoi-based space partitioning schemes as presented in this chapter might also be beneficial to represent a geospatial web service profile. Such profiles might be generated by different resources/peers by the invocation of the *getCapabilities()* method which is defined by many OGC standards such as Web Map Service (WMS; http://www.opengeospatial.org/standards/wms), Web Feature Service (WFS; http://www.opengeospatial.org/standards/wfs), or Web Coverage Service (WCS; http://www.opengeospatial.org/standards/wcs). Currently, for example in case of WCS and WFS, Voronoi-based coverage summaries are not intended.

Resource description and selection techniques might also be applied within sensor networks (Elahi, et al., 2009). In sensor networks, limited processing power, bandwidth, and energy capacities necessitate aggregation techniques which are based on local information with a clear focus on space efficiency. One might think of in-situ or remote sensing applications, which aggregate and summarize gathered data before transmitting them in the network.

Lupu et al. (2007) present an approach for information sharing in mobile ad hoc networks. When people meet at certain events or places only for a limited amount of time, there might not be enough time to transfer full index data when a person for example searches for media items in the proximity of a given query location.

Compact geospatial resource descriptions might also be valuable for focused crawling (cf. Ahlers & Boll, 2009). If a service provides summaries of the geospatial extend of a certain website or media archive, a crawler could estimate the potential usefulness of this resource for its focused crawling task before actually visiting the source. This way, crawl efficiency can be improved by preventing the crawler from analyzing too many irrelevant pages. Web traffic imposed by downloading large irrelevant data volumes can thus be avoided.

Distributed IR techniques can also be used for vertical selection within aggregated search (Arguello, et al., 2009). Vertical selection is the task of identifying relevant verticals, i.e. focused search services such as image, news, video or shopping search. A user issuing the textual query "Beatles Hamburg" might also be interested in music videos captured in Hamburg and thus the results of video search or small previews should be integrated in result presentation of classical web search if video files match the geospatial restriction. In this context, a vertical can be interpreted as a resource and the task of selecting relevant verticals is similar to resource selection in distributed IR requiring adequate features, i.e. resource descriptions, and corresponding selection mechanisms.

Space efficient geospatial resource descriptions might also be beneficial in the context of recommender systems and social search for example in order to compute the similarity between different users of social network sites. Similar users can be determined not only based on having the same friends, using the same tags, bookmarking the same media items, etc. (Guy, et al., 2010), but also depending on the similarity of geospatial footprints which are obtained from the media items a user administers.

For many scenarios, it might be necessary to enhance the basic description and selection techniques presented in this chapter in order to summarize for example trajectories or more complex objects such as polygons. Also in this context, the literature on multidimensional and metric index structures provides a good starting point for further studies (cf. Samet, 2006). Many multidimensional access methods are capable of administering lines and polygons, which are aggregated for example in the inner nodes of tree-based access methods in an adequate way.

FUTURE RESEARCH DIRECTIONS

So far, our heuristic approach lacks the availability of an adequate stopping criterion. In our experiments we analyze how many peers need to be contacted in order to retrieve the 20 closest locations according to a given query location. Of course, we could use these empirical findings in order to derive a mechanism which tries to guarantee that for example in 90% of the queries all 20 nearest neighbors can be found. Nevertheless, we doubt if such an approach adapts to different media types and collection sizes. Thus, it is necessary to design algorithms for k-NN query processing which can successfully prune peers from query processing if they do not contribute relevant documents.

Another interesting aspect might be to further optimize the GRID and BB approaches. For BB, techniques, which use more than one bounding box per resource, might be promising (cf. Becker, et al., 1991; Chen, et al., 2006). For the GRID approach, triangular partitions as well as a partitioning which results in cells of equal size might be considered (cf. Szalay, et al., 2005; Putman, et al., 2007).

It might also be interesting to further analyze ranking selectivity when very remote places on earth are used as query locations. If they represent places in certain countries with low GDP, this might lead to a loss in ranking selectivity for UFS and HFS. In these cases, other strategies for obtaining the centroids might be used. In addition, BB or GRID might be better suited in such cases.

As mentioned before, we also plan to use UFS and HFS to enhance centralized index structures.

In general, there is a need for resource description and selection techniques based on local image features. Here, image content is represented by several feature vectors per image. Local features are also used in video retrieval. A further research direction might be to apply HFS and UFS in this context.

Since resource selection techniques for text and media content information are available, one might take a closer look at combined queries. If a user queries for "Beatles Hamburg" it is not sufficient to rank peers high, which contain documents addressing the Beatles or the city of Hamburg. A user might expect documents which match both criteria.

CONCLUSION

In this chapter, we have outlined different P2P approaches for the large-scale administration of media content. We identified geospatial information as an important search criterion and discussed how it can be applied in the various P2P approaches. Later in the chapter, we presented and evaluated different resource description and selection techniques for geospatial k-NN queries and identified UFS as a promising approach. Here, binary histograms are used which capture if a certain reference location is closest to one or more of a resources media locations. We also outlined how UFS can be applied in a P2P scenario. In addition, we qualitatively argued that geospatial resource description and selection techniques such as UFS might also be promising in various other application fields.

REFERENCES

Ahlers, D., & Boll, S. (2009). Adaptive geospatially focused crawling. In D. Cheung, I. Song, W. Chu, X. Hu, J. Lin, J. Li, & Z. Peng (Eds.), *Proceedings of the 18th ACM International Conference on Information and Knowledge Management,* (pp. 445-454). Hong Kong, China: ACM.

Arguello, J., Diaz, F., Callan, J., & Crespo, J.-F. (2009). Sources of evidence for vertical selection. In M. Sanderson, C. Zhai, J. Zobel, J. Allan, & J.-A. Aslan (Eds.), *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 315-322). Boston, MA: ACM.

Beaver, D., Kumar, S., Li, H. C., Sobel, J., & Vajgel, P. (2010). Finding a needle in haystack: Facebook's photo storage. In *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation (OSDI 2010)*. Berkeley, CA: USENIX Association.

Becker, B., Franciosa, P. G., Gschwind, S., Ohler, T., Thiemt, G., & Widmayer, P. (1991). An optimal algorithm for approximating a set of rectangles by two minimum area rectangles. In H. Bieri & H. Noltemeier (Eds.), *Proceedings of the International Workshop on Computational Geometry*, (pp. 13-25). Berlin, Germany: Springer.

Belkin, N. J., Kantor, P., Fox, E. A., & Shaw, J. A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management, 31*(3), 431–448. doi:10.1016/0306-4573(94)00057-A

Bender, M., Michel, S., Weikum, G., & Zimmer, C. (2005). The minerva project: Database selection in the context of P2P search. In G. Vossen, F. Leymann, P. C. Lockemann, W. Stucky (Eds.), *Proceedings of Datenbanksysteme in Business, Technologie und Web, 11: Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme"*, (pp. 125-144). Karlsruhe, Germany: DIS.

Blank, D., El Allali, S., Müller, W., & Henrich, A. (2007). Sample-based creation of peer summaries for efficient similarity search in scalable peer-to-peer networks. In R. Lienhart, A. R. Prasad, A. Hanjalic, S. Choi, B. P. Bailey, & N. Sebe (Eds.), *Proceedings of the 15th International Conference on Multimedia*, (pp. 143-152). Augsburg, Germany: ACM.

Blank, D., & Henrich, A. (2009). Summarizing geo-referenced photo collections for image retrieval in P2P networks. In *Proceedings of the International Workshop on Geographic Information on the Internet*, (pp. 55-60). Retrieved from http://georama-project.labs.exalead.com / workshop/GIIW-proceedings.pdf.

Blank, D., & Henrich, A. (2010). Description and selection of media archives for geographic nearest neighbor queries in P2P networks. In A. R. Doherty, C. Gurrin, G. J. F. Jones, & A. F. Smeaton (Eds.), *Proceedings of Information Access for Personal Media Archives Workshop*, (pp. 22-29). Received from http://doras.dcu.ie/15373/.

Bockting, S., & Hiemstra, D. (2009). Collection selection with highly discriminative keys. In *Proceedings of the 7th International Workshop on Large-Scale Distributed Systems for Information Retrieval*. Retrieved from http://lsdsir09.isti.cnr.it/ lsdsir09-1.pdf.

Chen, S., Liang, S., & Wang, M. (2010). A locality-aware peer-to-peer approach for geospatial web services discovery. In *Canadian Geomatics Conference*. Retrieved from http://www.isprs.org /proceedings/XXXVIII/part1/13/ 13_01_Paper_65.pdf.

Chen, Y.-Y., Suel, T., & Markowetz, A. (2006). Efficient query processing in geographic web search engines. In S. Chaudhuri, V. Hristidis, & N. Polyzotis (Eds.), *Proceedings of the 25th International Conference on Management of Data*, (pp. 277-288). Chicago, IL: ACM.

Cuenca-Acuna, F., Peery, C., Martin, R. P., & Nguyen, T. D. (2003). PlanetP: Using gossiping to build content addressable peer-to-peer information sharing communities. In *Proceedings of IEEE International Symposium on High Performance Distributed Computing*, (pp. 236-246). Seattle, WA: IEEE Press.

Dolin, R., Agrawal, D., Abbadi, A. E., & Dillon, L. K. (1997). Pharos: A scalable distributed architecture for locating heterogeneous information sources. In F. Golshani & K. Makki (Eds.), *Proceedings of the 6th International Conference on Information and Knowledge Management,* (pp. 348-355). Las Vegas, Nevada: ACM.

Doulkeridis, C., Vlachou, A., Nørvåg, K., Kotidis, Y., & Vazirgiannis, M. (2009). Multidimensional routing indices for efficient distributed query processing. In D. Cheung, I. Song, W. Chu, X. Hu, J. Lin, J. Li, & Z. Peng (Eds.), *Proceedings of the 18th ACM International Conference on Information and Knowledge Management,* (pp. 1489-1492). Hong Kong, China: ACM Press.

Doulkeridis, C., Vlachou, A., Nørvåg, K., & Vazirgiannis, M. (2010). Distributed semantic overlay networks. In Shen, X., Yu, H., Buford, J., & Akon, M. (Eds.), *Handbook of Peer-to-Peer Networking.* Berlin, Germany: Springer. doi:10.1007/978-0-387-09751-0_17

Eisenhardt, M., Müller, W., Blank, D., El Allali, S., & Henrich, A. (2008). Clustering-based, load balanced source selection for CBIR in P2P networks. *International Journal of Semantic Computing, 2*(2), 235–252. doi:10.1142/S1793351X08000439

Elahi, B. M., Römer, K., Ostermaier, B., Fahrmair, M., & Kellerer, W. (2009). Sensor ranking: A primitive for efficient content-based sensor search. In *Proceedings of the 8th International Conference on Information Processing in Sensor Networks,* (pp. 217-228). Washington, DC: IEEE.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. In B. Yormark (Ed.), *Proceedings of ACM SIGMOD Conference,* (pp. 47-57). Boston, MA: ACM.

Guy, I., Jacovi, M., Perer, A., Ronen, I., & Uziel, E. (2010). Same places, same things, same people? Mining user similarity on social media. In *Proceedings of the 22nd International Conference on Computer Supported Cooperative Work,* (pp. 41-50). Savannah, GA: ACM.

Hariharan, R., Hore, B., & Mehrotra, S. (2008). Discovering GIS sources on the web using summaries. In R. Larsen, A. Paepcke, J. L. Borbinha, & M. Naaman (Eds.), *Proceedings of the 8th ACM/IEEE Joint Conference on Digital Libraries,* (pp. 94-103). Pittsburgh, PA: ACM Press.

Ilyas, I. F., Beskales, G., & Soliman, M. A. (2008). A survey of top-k query processing techniques in relational database systems. *ACM Computing Surveys, 40*(4), 1–58. doi:10.1145/1391729.1391730

Lupu, M., Li, J., Ooi, B. C., & Shi, S. (2007). Clustering wavelets to speed-up data dissemination in structured P2P manets. In *Proceedings of the 23rd International IEEE Conference on Data Engineering,* (pp. 386-395). Istanbul, Turkey: IEEE Press.

Memon, F., Tiebler, D., Dürr, F., Rothermel, K., Tomsu, M., & Domschitz, P. (2009). Scalable spatial information discovery over distributed hash tables. In *Proceedings of the Fourth International ICST Conference on COMmunication System softWAre and middlewaRE (COMSWARE 2009).* New York, NY: ACM.

Müller, W., Eisenhardt, M., & Henrich, A. (2005). Scalable summary based retrieval in P2P networks. In A. Chowdhury, N. Fuhr, M. Ronthaler, H.-J. Schek, & W. Teiken (Eds.), *Proceedings of the 14th ACM International Conference on Information and Knowledge Management,* (pp. 586-593). Bremen, Germany: ACM Press.

Nogueras-Iso, J., Zarazaga-Soria, F. J., & Muro-Medrano, P. R. (2005). *Geographic information metadata for spatial data infrastructures: Resources, interoperability and information retrieval.* Secaucus, NJ: Springer-Verlag.

Novak, D., Batko, M., & Zezula, P. (2008). Web-scale system for image similarity search: When the dreams are coming true. In *Proceedings of 6th International Workshop on Content-Based Multimedia Indexing,* (pp. 446-453). London, UK: IEEE.

Papapetrou, O., Siberski, W., Balke, W.-T., & Nejdl, W. (2007). DHTs over peer clusters for distributed information retrieval. In *Proceedings of the 21st International Conference on Advanced Information Networking and Applications,* (pp. 84-93). Niagara Falls, Canada: IEEE.

Putman, W. M., & Lin, S.-J. (2007). Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics, 227*(1), 55–78. doi:10.1016/j.jcp.2007.07.022

Ratnasamy, S., Francis, P., Handley, M., Karp, R., & Schenker, S. (2001). A scalable content-addressable network. In *Proceedings of the Annual Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication,* (pp. 161-172). San Diego, CA: ACM.

Samet, H. (2006). *Foundations of multidimensional and metric data structures.* San Francisco, CA: Morgan Kaufmann Publishers Inc.

Shen, X., Yu, H., Buford, J., & Akon, M. (2010). *Handbook of peer-to-peer networking.* Berlin, Germany: Springer. doi:10.1007/978-0-387-09751-0

Sinnott, R. (1984). Virtues of the haversine. *Sky and Telescope, 68*(2), 158.

Stoica, I., Morris, R., Karger, D., Kaashoek, F., & Balakrishnan, H. (2001). Chord: A scalable peer-to-peer lookup service for Internet applications. In *Proceedings of the Annual Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication,* (pp. 149-160). San Diego, CA: ACM.

Szalay, A., Gray, J., Fekete, G., Kunszt, P., Kukol, P., & Thakar, A. (2005). *Indexing the sphere with the hierarchical triangular mesh.* Technical Report: MSR-TR-2005-123. Retrieved from http://research.microsoft.com /pubs/64531/tr-2005-123.pdf.

Thomas, P., & Hawking, D. (2009). Server selection methods in personal metasearch: A comparative empirical study. *Information Retrieval, 12*(5), 581–604. doi:10.1007/s10791-009-9094-z

Vincenty, T. (1975). Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review, 22*(176), 88–93.

Vu, Q. H., Lupu, M., & Wu, S. (2009). Simpson: Efficient similarity search in metric spaces over P2P structured overlay networks. In H. Sips, D. Epema, & H.-X. Lin (Eds.), *Proceedings of the 15th International Euro-Par Conference on Parallel Processing,* (pp. 498-510). Delft, The Netherlands: Springer.

Xiujun, M., Gang, L., Kunqing, X., & Meng, S. (2006). A peer-to-peer approach to geospatial web services discovery. In *Proceedings of the 1st International Conference on Scalable Information Systems (InfoScale 2006).* New York, NY: ACM.

KEY TERMS AND DEFINITIONS

Peer-to-Peer (P2P) System: A P2P system is made up of distributed resources (i.e. computing devices) which cooperate in order to provide and consume certain services in a decentralized fash-

ion. In contrast to traditional client/server systems, a peer can act as both, a client and a server.

Peer-to-Peer Information Retrieval (P2P IR) System: A P2P system with a focus on the administration and retrieval of media items. Hereby, content-based information retrieval techniques are applied in order to search for media items.

Structured P2P IR Systems: P2P IR systems which are based on distributed index structures with distributed hash tables (DHTs) being the most prominent class member. In structured P2P IR systems, every peer is responsible for a certain region of the feature space.

Unstructured P2P IR Systems: In contrast to structured P2P IR systems, connections in unstructured P2P IR overlays do not emerge from a distributed index structure and are thus formed more arbitrarily. Here, every peer usually administers the index data of its own media objects.

Resource Description: An important task in distributed IR which provides the basis for query routing. The resources have to describe their content in an adequate way—optimizing the trade-off between selectivity and space efficiency of the data summaries (also called resource descriptions).

Resource Selection: The process of determining the order in which distributed resources should be contacted to fulfill a certain information need. Besides knowing the order in which resources should be contacted it is also important to determine when it is no longer beneficial to contact further resources.

Summary: The term summary also refers to resource descriptions. It emphasizes the need for space efficiency of the resource descriptions in certain application contexts, e.g. unstructured P2P IR systems.