Description and Selection of Media Archives for Geographic Nearest Neighbor Queries in P2P Networks

Daniel Blank
University of Bamberg
D-96052 Bamberg, Germany
daniel.blank@uni-bamberg.de

Andreas Henrich
University of Bamberg
D-96052 Bamberg, Germany
andreas.henrich@uni-bamberg.de

ABSTRACT

In recent years, there has been a tremendous increase in personal media data stored on people's PCs, mobile devices, and social media sites in the web. Additionally, people are increasingly collaborating and interacting by sharing and commenting media items. These trends call for retrieval services integrating resources heterogeneous in update frequencies, media types, and size. In this context, peer-to-peer (P2P) technologies offer interesting solutions. When performing a query on certain types of P2P networks, resource selection is important. Compact summaries (i.e. resource descriptions) of each peer's data collection are known to other peers and used by them in order to determine promising peers for a given query. Summaries have to describe not only textual information, content-based media features, and information about date and time, but also the locations where e.g. images were taken, videos were recorded, or to which a user travelled. The present paper proposes and evaluates different resource selection techniques based on descriptions of the geographic footprint of personal media archives when querying for media items that are geographically close to a given query location. These techniques are not restricted to P2P networks and can e.g. be applied in hybrid index structures or distributed IR systems in general.

Categories and Subject Descriptors: H3.3 Information Storage and Retrieval: Information Search and Retrieval [Search process, Selection process]; H3.4 Information Storage and Retrieval: Systems and Software [Distributed systems, Performance evaluation]

General Terms: Algorithms, Performance

Keywords: Resource Selection, Peer-to-Peer IR, Multimedia IR, Summarization

1. INTRODUCTION

There has been a tremendous increase in personal media data during the last years. People write blogs, twitter about their lives, and use remote photo and video communities. Besides customized services such as Flickr or Youtube, cloud services offer more general storage facilities (e.g. `https://one.ubuntu.com/`, all URLs last visited on 13.1.10). In addition, personal information management (PIM) becomes important to administer e-mail accounts, personal contacts, appointments, tasks, and notes accessible from multiple devices. Consequently, heterogeneous online resources differing in size, media type, and update characteristics have to be administered [18]. Besides storing personal media items, people tend to share these items with friends and interact with each other by collaboratively tagging or commenting on various items.

In our scenario, personal media archives are administered in a P2P system. All media items of an archive are stored locally on the peer's/user's personal device without a need to store items on remote servers hosted by service providers such as Flickr or Youtube. Besides avoiding dependance on service providers as informational gatekeepers, no expensive infrastructure has to be maintained by applying a scalable P2P protocol such as Rumorama [20]. Idle computing power in times of inactivity can be used to maintain, analyze and enrich media items. For the purpose of sharing and collaboration, friends or groups can be granted access to certain media items of a peer/user.

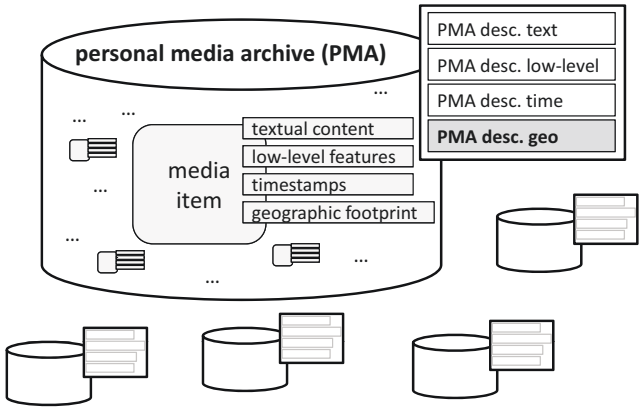

Figure 1: Criteria for media archive selection.

In order to facilitate retrieval, media items are described by four criteria: 1) textual content, 2) low-level features, 3) timestamps, and 4) a geographic footprint. Personal media archives containing multiple media items can thus be represented by four corresponding summaries (cf. Fig. 1) allowing for efficient and effective archive selection when processing

IAPMA'10, March 28, Milton Keynes, UK

a given query. In literature many approaches for P2P information retrieval (IR) can be found (cf. Sect. 2). However, these approaches mostly do not consider a unified search scenario combining all criteria from Fig. 1. To our best knowledge, the approach outlined here is the first with this target in mind. We analyze geographic resource description and selection techniques in this paper (highlighted in Fig. 1). Different geo-summaries are evaluated w.r.t. geographic nearest neighbor queries in Sect. 4. Techniques for summarizing textual and media content information in form of high-dimensional, real-valued feature vectors have already been proposed (cf. [6, 11] for example). Summarization of time and date information seems possible by a combination of clustering and histogram techniques (cf. Sect. 2). Of course, resource ranking based on a single criterion is only a first step. When querying for multiple criteria, individual resource rankings can be merged by applying a merging algorithm for ranked lists (cf. [3]).

Our approach is based on, but not restricted to, Rumorama [20], a scalable P2P protocol building hierarchies of PlanetP-like networks accessible by an efficient multicast. In PlanetP [11], randomized rumor spreading assures that every peer knows summaries of all other peers in the network. Summarizations of other peers' data provide the basis for query routing decisions, i.e. which peers to contact during query processing. While we examine peer summaries for geographic data in PlanetP-like middle-sized networks, results can be applied within large-scale Rumorama-like P2P networks. We also believe that results can be transferred to other application domains (cf. Sect. 2). Additionally, summarizations can be visualized (cf. Fig. 2 and 4). This might be beneficial for interactive retrieval e.g. by providing a visual overview of personal media archives for a huge number of archives with low bandwidth requirements.

The paper is organized as follows. Section 2 gives a brief overview on related work. In Section 3 we describe the resource description and selection techniques we employ. The experiments are part of Section 4. Here, we use two collections of geotagged images with user-information which allows for a realistic, user-oriented distribution of images to resources. Section 5 consists of a conclusion and an outlook on future work.

2. RELATED WORK

P2P IR systems can be classified into several groups. Systems such as PlanetP [11] and Rumorama [20] follow a semantic query routing approach based on resource descriptions generated and transmitted by every peer.

The second group of approaches are semantic overlay networks (e.g. [14]) where the content of the peer's data defines its place in the network topology. Peers are organized by semantic clusters and within query execution the query has to be forwarded to the most promising clusters. Here, the simultaneous indexing of multiple criteria as depicted in Fig. 1 would require the definition of a similarity between peers and images combining e.g. geographic and image content information. Alternatively, multiple overlays might be maintained.

A third class of P2P systems are structured networks such as distributed hashtables (DHTs). Novak et al. [7] present a large-scale architecture based on DHTs. Within DHTs, indexing data is transferred to remote peers with every peer being responsible for a certain range of the feature space. Presumably, correlations between different criteria (e.g. geographic and image content information) are difficult to exploit. If we e.g. assume an image from the Sahara Desert with shades of beige sand and blue sky, different peers might be responsible for indexing the geographic and the image content information. Therefore, when distributing the indexing data of the Sahara image, querying for it, or removing it from the network, (at least) two different peers have to be contacted. In the case of high-dimensional feature vectors, a problem within structured P2P networks is the load imposed onto the network when inserting new documents. If many peers do this at the same time, the insertion of data might become a bottleneck. Therefore, both, super-peer as well as summarization techniques are used within DHTs. While super-peer approaches are out of the scope of this paper, the strategy of creating peer summaries that are later indexed in a DHT is e.g. applied in [15] for content-based image features. In general, there is a convergence of structured and unstructured P2P networks with many hybrid approaches. DHTs have also been applied for social (semantic) desktops, e.g. in the Nepomuk project (`http://nepomuk.semanticdesktop.org`), which supports collaborative PIM and the sharing of media items. Nepomuk allows full-text search in combination with RDF-based queries. We see our work complementary in two directions. First, it tries to overcome some limitations of DHTs. Second, it allows for content-based multimedia retrieval enhanced with geographic nearest neighbor queries.

There is plenty of work on resource selection in traditional, distributed IR—especially for text data (for references see e.g. [18]). Most of the proposed resource descriptions are designed to maximize retrieval quality within the context of server selection. Within such a client/server scenario, space efficiency of resource descriptions is less important. Hence, summaries in distributed IR are usually less space efficient than the ones we are looking at.

Clustering (cf. [9]) might be used to represent collections of geo-locations by their centroids. Recently, cluster hulls have been proposed that summarize sets of geo-locations by several convex hulls [13]. We expect the summaries generated by these techniques to be less space efficient than the most promising approaches from Section 3. Here, image locations are assigned to the closest reference vector, which can be interpreted as a special form of clustering, leading to space efficient resource descriptions.

Tree-based index structures are also related to this work (cf. R-tree and variants [12, 16]). The decision of choosing the best subtree is similar to the peer selection problem. Summaries in the P2P context correspond to summaries maintained in the nodes of a tree (e.g. bounding boxes). Becker et al. [2] present an algorithm for summarizing a set of bounding boxes by two bounding boxes minimizing the area that is covered. Chen et al. [5] 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 that is not covered by any indexing data. These approaches demonstrate that there is optimization potential with the representation of geo-regions in spatial access structures. However the mentioned approaches stick with bounding boxes and the approaches presented in this paper might be interesting alternatives in this application field. A detailled comparison will be part of future work.

Our summarization techniques might also be applied

within sensor [10] as well as ad hoc networks [15]. Because of limited processing power, bandwidth and energy capacities it is essential that aggregation techniques used within sensor networks are based on local information with a clear focus on space efficiency. Lupu et al. [15] present an interesting approach for ad hoc information sharing based on mobile devices when people meet at certain events or places. Here, it might not be feasible to share the complete data but only summarized information.

Compact resource descriptions might also be valuable for geographically focused crawling [1]. If a service provides summaries of the geographic extend covered by 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. Geoparsing of web pages as well as downloading large sets of images in order to extract EXIF information can be avoided.

3. RESOURCE DESCRIPTION & SELECTION FOR GEOGRAPHIC QUERIES

The description and selection of resources is common in distributed IR (cf. Sect. 2). In general, there is a trade-off between the quality of resource descriptions and their size. Larger summaries can encode more information and should therefore allow for better resource selection. Within most of the distributed IR literature there is a clear focus on the quality of the resource descriptions. Often, only few resources with static data are analyzed making space efficiency of the resource descriptions not the main optimization target. We, on the contrary, need to find a more balanced solution. For our scenario it is essential to apply summaries that are space efficient and at the same time meaningful and selective enough to allow for efficient and effective resource selection, mainly because of two reasons. First, PlanetP-like networks might consist of more resources than many other systems in distributed IR. Second, resources might (re)join the network and the content administered by the resources might change frequently triggering new resource descriptions to be generated and sent.

Resource selection is performed by ranking the peers based on a set of resource descriptions, the query location and maybe some additional information such as reference points. The peer ranking defines the order in which peers are being contacted during query processing. When searching for the k closest images w.r.t. a query location, the peer ranking should reflect that peers administering a bigger fraction out of the top-k images receive a higher rank than peers maintaining a smaller fraction of top-k images.

In the following we will present four different resource description and corresponding selection techniques. The selectivity of three of the resource description techniques is briefly analyzed in [4]. Query processing is not discussed in [4]. Thus, the design and analysis of different resource selection techniques for k nearest neighbor queries and the improvement of the most promising resource description techniques are the main contribution of this paper.

Bounding Boxes (BB)

When using BB as resource description every peer computes a bounding box over the geographic coordinates of its image collection (cf. Fig. 2, left). We encode a latitude/longitude-pair (for short: lat/lon-pair) with 8 bytes, 4 for latitude and 4 for longitude. Therefore, we require $8*2$ 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 within 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 within the bounding box of both peers p_a and p_b, the size of a peer (i.e. the number of images a peer administers) is used as an additional criterion. Peers with more images are ranked higher than peers with fewer images. If neither peer p_a nor peer p_b contains the query within its resource description, 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 $6,371$ km. If not stated otherwise, we use Haversine distance [17] to compute the distance between two points on the sphere.

Grid-based Summaries ($GRID_r$)

In a second approach, the geographic coordinate space is represented as a grid (cf. Fig. 2, middle). A parameter r is used to define 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 $\frac{180°}{r} = \frac{360°}{2r}$ in the latitude and longitude domain. This simplified view is e.g. also applied in [8] 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. Sect. 4.4). 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. Within the summary, bit positions are determined horizontally from left to right and from bottom to top. Effects of alternative strategies on compression will be evaluated in future work.

During peer ranking the grid cell containing the query location is determined. If peer p_a has an image within 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, GRID 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[1] based on $1 + 8 = 9$ cells and in the third round on $1 + 8 + 16 = 25$ cells and so on. The ranking criterion in every round is the number of grid cells containing one or more image location(s) (the more the better).

Highly Fine-grained Summaries (HFS_n)

In this case we use resource descriptions originally designed for summarizing image content based on the color distribution or texture of an image [6]. We randomly choose a set of n image locations from the global image collection as

[1]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-degree meridian we assume that there is no boundary and neighborhood relations are valid in both directions.

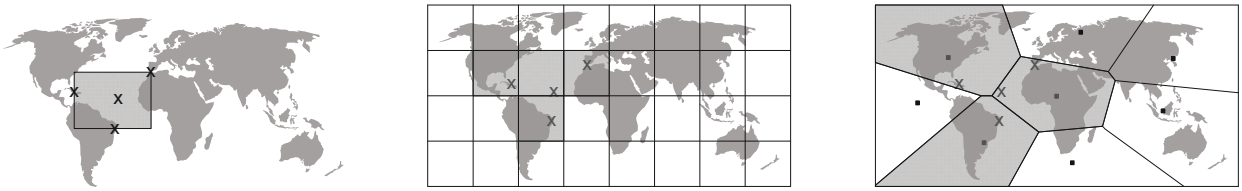

Figure 2: Visualizing summary creation for *BB* (left), $GRID_4$ (middle) and HFS_8/UFS_8 (right: ■ corresponds to reference points). Four images are geotagged, indicated as x.

reference points. This set of reference points is known to all peers[2]. Every image of a peer's local image collection is afterwards assigned to the closest reference point according to Haversine distance (cf. Fig. 2, 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 centroid c_j $(1 \leq j \leq n)$.

Peer ranking is performed as follows. The reference points c_j are sorted in ascending order according to Haversine distance w.r.t. the query. The first element of the sorted list L corresponds to the cluster centroid 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 recursively ranked w.r.t. the number of documents within this cluster.

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 centroid j is the closest centroid 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 centroids being used which might result in similar or even improved retrieval performance compared to HFS. Among other aspects, this will be evaluated in Section 4.4.

4. EVALUATION

In the following we will present the data collections we use (Sect. 4.1), some rough calculations and basic considerations justifying the use of summarized geographic resource descriptions (Sect. 4.2), main characteristics of the experimental setting (Sect. 4.3) and the experiments themselves (Sect. 4.4).

4.1 Data Collections

We use two data collections of geotagged images:

Geoflickr: During 2007 we crawled a large amount of publicly available images which had been uploaded to Flickr. 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 geotagged. After some data cleansing the Geoflickr collection consists of 406,450 geotagged images from 5,951 different users/peers.

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 geotagged images and also distributed them to peers in a user-centric approach. In our scenario every Geograph participant operates a single peer; 2,609 peers administer 246,937 images in total.

The distribution of the number of images per peer is displayed for both collections in Figure 3. For both collections the distribution of the number of images per peer is skew which is typical for P2P networks [11]. Approximately the first 1% of the biggest peers per collection, i.e. the 60 biggest peers for the Geoflickr collection and 26 biggest peers for the Geograph collection, administer 42.0% and 28.3% of the collection's images respectively. Whereas the biggest peer of the Geoflickr collection maintains 8.8% of the images, the biggest peer of the Geograph collection maintains 3.5%. In opposition, approximately 20.7% and 17.7% of the peers administer only a single image for the Geoflickr and the Geograph collection respectively. Approximately 50% of the images are maintained by 1.8% resp. 2.7% of the peers for the Geoflickr and the Geograph collection.

Figure 4 shows the geographic distribution of image locations. The Geoflickr collection consists of photos taken in various parts of the world with a focus on North America, Europe and Japan. In contrast, images of the Geograph collection are limited to the UK and Ireland with images more densily located around urban areas such as e.g. London.

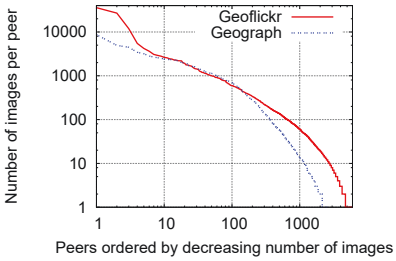

Figure 3: Number of images per peer for *Geoflickr* (5,951 peers) and *Geograph* (2,609 peers) collection.

4.2 Estimating Data Transfer Volume

The use of resource description and selection techniques is justifiable by several reasons. First, there might be scenarios e.g. in the context of ad hoc or sensor networks where it is not feasible or desirable to transfer complete indexing

[2]For CBIR, obtaining reference points from an external source and distributing them with software updates is proposed in [6]. We believe that this strategy is directly applicable also for geographic data. Strategies for selecting the reference points are evaluated in Sect. 4.4. Some peers could monitor the distribution of image locations in order to select appropriate reference points.

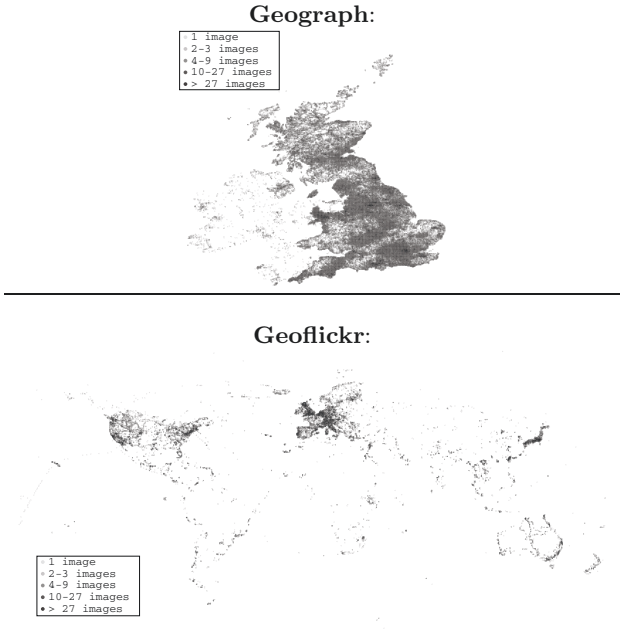

Figure 4: Geographic distribution of image locations for *Geograph* (top) and *Geoflickr* (bottom).

or sensor data (cf. Sect. 2). Second, resource descriptions might be used as "aggregators" to boost efficiency of certain types of applications such as focused crawling (cf. Sect. 2). Third, in certain scenarios the use of summaries will be beneficial compared to the transfer of full indexing data. Here, the total cost[3] of indexing (C^s_{idx}) and querying (C^s_q) information with the use of summarization techniques is smaller than the cost of transferring unsummarized indexing data C^u_{idx}. In the latter case resource selection becomes trivial, because the complete query processing can be performed locally ($C^u_q = 0$). So, resource selection is beneficial in terms of network traffic as long as:

$$C^s_{idx} + C^s_q < C^u_{idx}$$

We assume a message overhead m_o for header information, peer ID, timestamp, etc. and consider P resources and I documents/images. Furthermore, m_{idx} denotes the size of an index entry (e.g. 8 bytes in case of a lat/lon-pair) and m_s captures the average size of a resource description. Thus, in a non-dynamic network where resources stay online all the time and do not change their content, the cost of indexing is denoted as:

$$C^s_{idx} = (m_o + m_s) * P * (P-1)$$
$$C^u_{idx} = m_o * P * (P-1) + m_{idx} * I * (P-1)$$

It follows that resource selection is beneficial as long as:

$$C^s_q < (P-1) * (m_{idx} * I - m_s * P)$$

Thus, C^s_q has e.g. to be smaller than 15.8 GB and 4.5 GB for Geoflickr and Geograph respectively assuming average summary sizes of $m_s = 100$ bytes (cf. Sect. 4.4).

[3]In a rough calculation cost is measured in terms of data volume that is transferred and ignoring the storage requirements for replicating the indexing data.

An upper bound for the cost of querying summary-based networks can be derived depending on the average number of contacted peers $f * P$ ($0 \leq f \leq 1$), the cost for sending the query to a resource ($m_o + m_{idx}$) and the cost for receiving the result set ($m_o + m_{idx} * k$) where we assume that in the worst case w.r.t. C^s_q all inquired peers send k result items:

$$C^s_q = Q * f * P * (2m_o + m_{idx} * (k+1))$$

For example, with UFS_{8192} on average 0.2% resp. 0.5% of peers are contacted for Geoflickr and Geograph (cf. Sect. 4.4). If we assume a message overhead of approximately 50 byte and Q queries in total, C^s_q becomes 3.2 $kB*Q$ and 1.4 $kB*Q$ for Geoflickr and Geograph indicating that under these conditions roughly 5.0 million and 1.3 million queries can be performed respectively, before exceeding the data volume C^u_{idx}.

Obviously, overall network load in a PlanetP-like setting also depends on the frequencies of peers (re)joining the network and the characteristics of peers updating their document collections. In general, the decision to use aggregated resource descriptions rather than the data itself has to be based on various network characteristics influenced by e.g. the application scenario and the type(s) of data being indexed.

Of course, there are further optimizations to our approach. For example, peers maintaining only few images might transfer the geo-locations directly whereas only big peers send aggregated resource descriptions. Such a hybrid resource description strategy affords hybrid peer ranking schemes which are not the scope of this work.

4.3 Experimental Setting

For every parameter combination we run at least 5,000 queries. If a randomized selection of reference points is needed we run 50 experiments with 100 queries each, which also results in 5,000 queries per parameter setting.

Space efficiency of different resource descriptions is measured by analyzing summary sizes (cf. Sect. 4.4). For compressing the summaries we apply Java's `gzip` implementation with default parameter values. Our measurements include serialization overhead necessary in order to distribute the resource descriptions within the network.

We use two modes for selecting a geo-location as query. First, we randomly choose a geo-location of an image from the entire document collection (*queryMode=1*). 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. Second, we select a random peer and from this peer we choose a geo-location of an image at random (*queryMode=2*). Here, it is more likely that also a small peer contributes to the top-k query results since peers are chosen equiprobable.

When measuring retrieval performance we determine the fraction of peers that needs to be contacted on average in order to retrieve a certain fraction of the top-k image locations ($k = 20$) w.r.t. a given lat/lon-pair as query location. The top-k geo-locations are computed using Vincenty distance [19]. Since we are interested in the quality of the resource 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 dis-

tance 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.).

4.4 Experiments

At first, we evaluate retrieval performance of the BB approach (cf. Sect. 3). Figure 5 (left) on the next page shows the fraction of peers contacted on average in order to retrieve a certain fraction of the 20 closest image locations to a given query location. For both collections it seems reasonable, in the case where the bounding box of both peers p_a and p_b contains the query location, to contact the peer that administers the larger number of images in total (*bc:size*). For *bc:bbsize* and *bc:bbrecipsize*, in the case when both bounding boxes of peers p_a and p_b contain the query location, an approximated size and the reciprocal of the approximated size of the bounding box is used respectively in order to make a peer ranking decision; *bc:bbsize* ranks a peer with larger approximated size of the bounding box higher while *bc:bbrecipsize* prefers a peer with smaller approximated size of the bounding box. For reasons of comparison, Fig. 5 (left) also includes a ranking solely based on the number of images a peer maintains (*size*). In this case, a peer is ranked higher the more images it administers.

For the Geoflickr collection *bc:bbrecipsize* performs similar to *bc:size* while preferring smaller peers since in general the approximated size of their bounding box is smaller compared to big peers. When solely ranking by peer *size* some big peers are contacted at an early stage that cannot contribute to the top-20.

It can be observed that for the Geograph collection the performance gap between *bc:size* on the one hand and *bc:bbsize* as well as *bc:bbrecipsize* on the other is more distinct compared to Geoflickr at least w.r.t. the retrieval of e.g. 80% of the top-k images, because in this case neither the approximated size of the bounding box nor the reciprocal are an adequate indicator for estimating the true *size* of a peer. In general, a ranking solely by peer *size* is less efficient for Geograph than for Geoflickr when querying for all of the top-20 image locations.

The size of a summary in the case of BB is 16 bytes for the bounding box plus 27 bytes serialization overhead, so 43 bytes in total. As it is desirable to include the size of a peer within the peer ranking process (*bc:size*), we assume additional 2 bytes for peer size information which leads to overall summary sizes of 45 bytes for BB (cf. Fig. 6).

BB performs better for the Geoflickr than for the Geograph collection which is likely to be due to the fact that for the latter there is more overlap among the bounding boxes [4]. Compared to the resource selection techniques presented in the following, BB performs worse. More peers are contacted on average in order to retrieve a certain fraction of the top-k query locations.

Retrieval performance of HFS/UFS is depicted in Fig. 5 (middle). As expected, differences in retrieval performance between HFS and UFS diminish with increasing number of summary bins 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.

Using UFS seems worthwhile if in addition to retrieval performance also overall summary sizes are analyzed (cf. Fig. 6). Average, minimum and maximum summary sizes are reduced when using UFS instead of HFS (cf. Tab. 1) indicating that both, resource descriptions of small as well as big peers can be represented in a more space efficient way. BB results in the most space efficient summaries, but retrieval performance is worse compared to HFS/UFS and GRID. Nevertheless, the use of BB might trade off when the network is not queried very frequently, which is a rather unrealistic assumption for P2P networks. GRID can be summarized more space efficiently than HFS and UFS, but retrieval performance is worse as will be shown in the following.

Results for GRID on both collections are shown in Fig. 5 (right). Obviously, retrieval performance is improved as more grid cells are used. For the Geograph collection, the grid is not adapted to the boundaries of the UK. Thus, retrieval performance is worse compared to the Geoflickr collection. Nevertheless, in the case where image locations are limited to a certain region of the world, an adaptation of the grid is possible. We did not adapt it in order to show the effects of a skew distribution of geographic image locations on a global scale. HFS and UFS are better suited for such scenarios than GRID, because they better adapt to the data that is used. Retrieval performance of HFS/UFS is also better compared to GRID when applied on a global scale, i.e. on the Geoflickr collection. Both, HFS_{8192}/UFS_{8192} as well as GRID_{64} result in the same number of summary bins in the uncompressed case ($64 * 2 * 64 = 8192$). Assuming *queryMode=2* and the Geoflickr collection, in the case of HFS/UFS 0.2% of peers are contacted in order to retrieve all top-20 image locations, whereas in the case of GRID 2.3% of the peers are contacted on average.

So far, we have assumed that the n reference points are chosen from the underlying data collection. Although this approach is feasible in general, we will now evaluate different

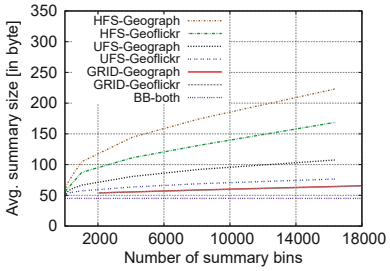

Figure 6: Avg. summary sizes, `gzip` applied to all but *BB*.

	Geoflickr HFS_n			Geograph HFS_n		
n	avg	min	max	avg	min	max
1,024	87.4	75.1	349.9	104.7	75.5	808.3
4,096	111.2	88.1	937.2	144.6	88.7	1,810.7
8,192	131.6	103.6	1,482.6	174.8	105.0	2,643.7
16,384	168.7	135.9	2,278.7	223.1	137.6	3,778.9
	Geoflickr UFS_n			Geograph UFS_n		
n	avg	min	max	avg	min	max
1,024	57.3	48.0	149.2	66.5	48.0	178.5
4,096	63.6	48.0	296.9	80.8	48.5	491.8
8,192	68.9	48.1	467.3	92.2	48.9	799.6
16,384	76.5	48.0	694.8	107.5	49.6	1,269.4

Table 1: Summary sizes for *HFS/UFS*

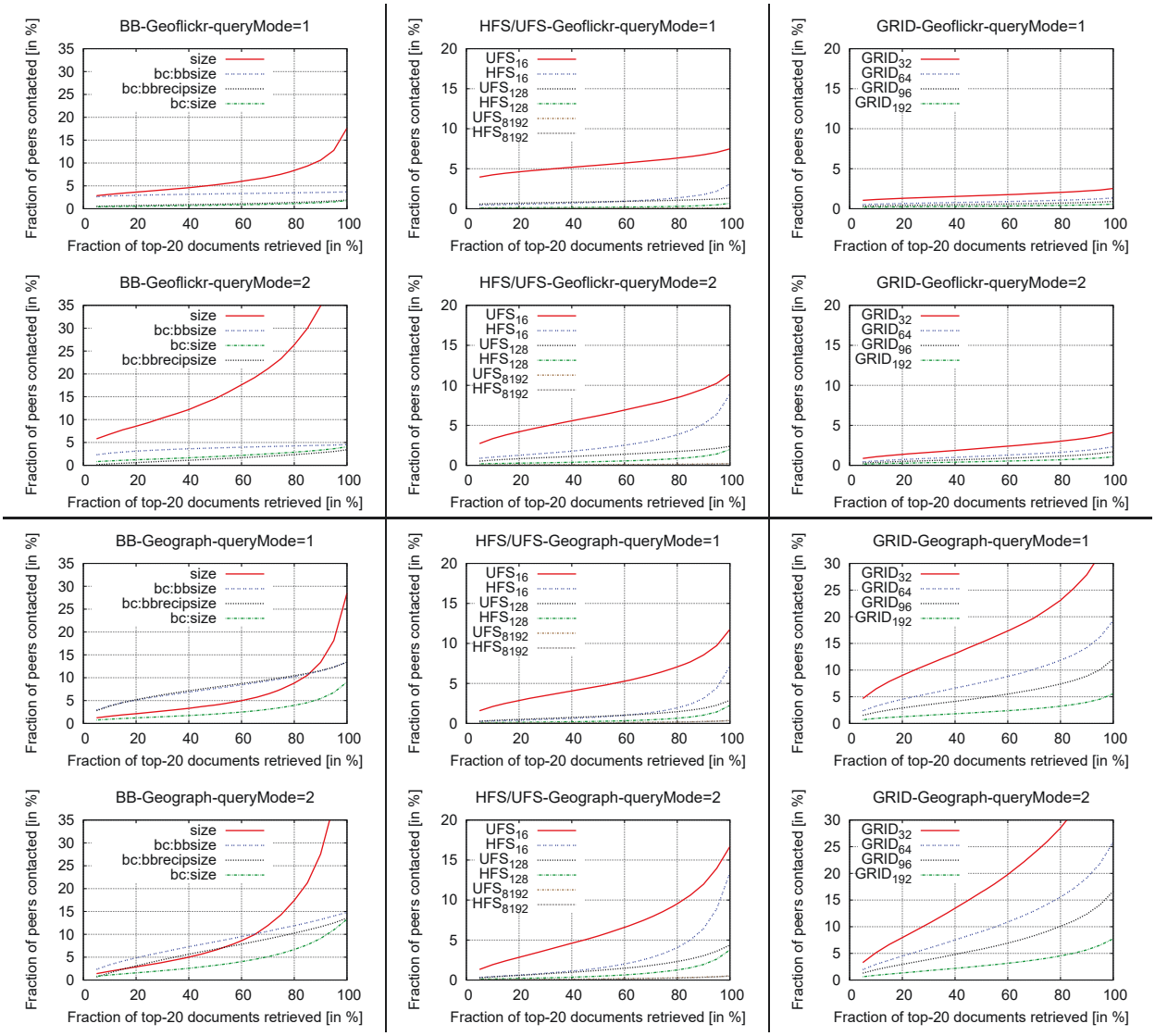

Figure 5: Fraction of peers contacted for *BB* (left), *HFS/UFS* (middle) and *GRID* (right)

sources for the reference points since we plan to distribute them with software updates from time to time in order to reduce overall network load.

We employ per country statistics from Worldmapper (`http://www.worldmapper.org/`) about mens' income (*INC*), Gross Domestic Product (*GDP*), population (*POP*), WWW usage (*WWW*) and land area (*LAND*). Based on these statistics we proportionally select the number of reference points from a certain country using Geonames gazetteer (`http://www.geonames.org/`). Reference points are selected amongst all populated places of a certain country at random. So, for example, if x% of the world's total mens' income is earned in a certain country, x% of the reference points are randomly chosen amongst all populated places of the specific country according to the information given by the Geonames gazetteer. For comparison we also choose reference points randomly distributed on the sphere (*RAND*).

Additional information is used from World Gazetteer (`http://world-gazetteer.com/`). We extract lists containing the geo-locations of the biggest cities in the *UK*, Europe (*EUR*) and the World (*GLOB*) in terms of population. A set of n reference points is chosen from each of these lists containing the n geographic locations with the most inhabitants according to the specific region.

Figure 7 shows results for *queryMode=2*. For reasons of brevity, we do not include figures for *queryMode=1* since in general they offer better retrieval performance as noticed before (cf. Fig. 5) and the same relative behavior as depicted for different strategies in Fig. 7. *COLL* represents the strategy where reference points are not determined from external sources but randomly chosen from the underlying data collection. On a global scale, when using external sources, selecting the centroids according to GDP is the most promising approach with performance similar to WWW and INC. These techniques adapt best to the data collections that are used and perform better than selecting e.g. the biggest cities in the world (GLOB). For both collections retrieval performance can be improved by increasing k e.g. up to $8,192$ or even higher. In general, Fig. 7 shows that *average* retrieval performance can be optimized by adequately selecting the

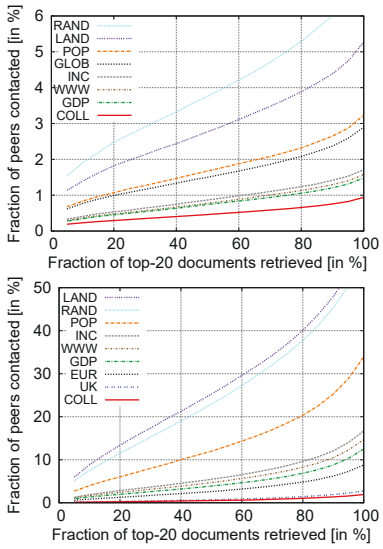

Figure 7: Sources of centroids for *UFS*$_{512}$; *Geoflickr* (top) and *Geograph* (bottom)

centroids according to the expected origin of query locations as well as image locations that are administered. This can e.g. be done by the implantation of special peers that analyze queries and the distribution of image locations. In Fig. 7 (bottom) the selection of centroids adapts to the collection by using the 512 biggest cities in the UK.

5. CONCLUSION & OUTLOOK

We have motivated a P2P system based on the description and selection of personal media archives. Media items are described by textual content, low-level multimedia features, timestamps, and geographic footprints. The focus of this paper is resource selection based on geographic information. In our experiments, both, bounding boxes as well as grid-based representations are outperformed in terms of retrieval performance by an approach using a set of reference points in order to compute (binary) histograms. Grid-based summaries show a higher potential for compression which might justify their use as well—depending on the characteristics of the usage scenario. Bounding boxes might be too coarse and induce overlap amongst peer summaries making selective peer ranking decisions difficult. Future work will mainly address the adaptation of UFS within CBIR as well as stopping criteria in order to determine when it is no longer beneficial to contact further peers during query processing.

Acknowledgement: Our work is partially funded by the Deutsche Forschungsgemeinschaft DFG HE 2555/12-2. We acknowledge the Geograph contributors (`http://www.geograph.org.uk/credits`). Their work is available under a Creative Commons Attribution-ShareAlike 2.0 Licence: `http://creativecommons.org/licenses/by-sa/2.0/`.

6. REFERENCES

[1] D. Ahlers and S. Boll. Adaptive geospatially focused crawling. In *Int. Conf. on Information and Knowledge Management*, pages 445–454, Hong Kong, China, 2009. ACM.

[2] B. Becker, P. G. Franciosa, S. Gschwind, T. Ohler, G. Thiemt, and P. Widmayer. An optimal algorithm for approximating a set of rectangles by two minimum area rectangles. In *Int. Workshop on Computational Geometry*, volume 553 of *LNCS*, pages 13–25. Springer, 1991.

[3] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Inf. Process. Manage.*, 31(3):431–448, 1995.

[4] D. Blank and A. Henrich. Summarizing georeferenced photo collections for image retrieval in P2P networks. In *Proc. of Workshop on Geographic Information on the Internet*, pages 55–60, `http://georama-project.labs.exalead.com/workshop/GIIW-proceedings.pdf` (12.11.2009), Toulouse, France, 2009.

[5] Y.-Y. Chen, T. Suel, and A. Markowetz. Efficient query processing in geographic web search engines. In *SIGMOD Conf.*, pages 277–288, Chicago, IL, USA, 2006. ACM.

[6] D. Blank, S. El Allali, W. Müller, A. Henrich. Sample-based creation of peer summaries for efficient similarity search in scalable peer-to-peer networks. *ACM SIGMM Workshop on Multimedia Information Retrieval (MIR 2007), Augsburg, Germany*, pages 143–152, 2007.

[7] D. Novak, M. Batko, P. Zezula. Web-scale system for image similarity search: When the dreams are coming true. In *Int. Workshop on Content-Based Multimedia Indexing*, pages 446–453, London, UK, 2008. IEEE.

[8] R. Dolin, D. Agrawal, A. E. Abbadi, and L. K. Dillon. Pharos: A scalable distributed architecture for locating heterogeneous information sources. In *Int. Conf. on Information and Knowledge Management*, pages 348–355, Las Vegas, Nevada, 1997. ACM.

[9] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2 edition, Nov. 2000.

[10] B. M. Elahi, K. Römer, B. Ostermaier, M. Fahrmair, and W. Kellerer. Sensor ranking: A primitive for efficient content-based sensor search. In *Int. Conf. on Information Processing in Sensor Networks*, pages 217–228, Washington, DC, USA, 2009. IEEE.

[11] F. Cuenca-Acuna and C. Peery and R.P. Martin and T.D. Nguyen. PlanetP: Using gossiping to build content addressable peer-to-peer information sharing communities. In *IEEE Int. Symp. on High Performance Distributed Computing*, pages 236–246, Seattle, WA, USA, 2003.

[12] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *SIGMOD Conf.*, pages 47–57, Boston, MA, 1984. ACM.

[13] J. Hershberger and N. Shrivastava and S. Suri. Summarizing spatial data streams using clusterhulls. *ACM Journal of Experimental Algorithmics*, 13:2.4–2.28, 2008.

[14] I. King, C. H. Ng, and K. C. Sia. Distributed content-based visual information retrieval system on peer-to-peer networks. *ACM Trans. Inf. Syst.*, 22(3):477–501, 2004.

[15] M. Lupu, J. Li, B. C. Ooi, and S. Shi. Clustering wavelets to speed-up data dissemination in structured P2P manets. In *ICDE*, pages 386–395, Istanbul, Turkey, 2007. IEEE.

[16] H. Samet. *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2006.

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

[18] P. Thomas and D. Hawking. Server selection methods in personal metasearch: a comparative empirical study. *Information Retrieval*, 12(5):581–604, 2009.

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

[20] W. Müller and M. Eisenhardt and A. Henrich. Scalable summary based retrieval in P2P networks. In *Int. Conf. on Information and Knowledge Management*, pages 586–593, Bremen, Germany, 2005. ACM.