Binary Histograms for Resource Selection in Peer-to-Peer Media Retrieval

Daniel Blank and Andreas Henrich
University of Bamberg
D-96052, Bamberg, Germany
{daniel.blank | andreas.henrich}@uni-bamberg.de

Abstract

With the ever increasing amount of media data and collections on the world wide web and on private devices arises 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 peer-to-peer (P2P) systems attractive. Here, resource selection can be applied to determine a ranking of promising resources based on descriptions of their content. Resources are contacted in ranked order to retrieve appropriate media items w.r.t. a user's information need.

In this paper we apply and adapt resource descriptions in the form of binary histograms and corresponding selection techniques which were designed for low-dimensional spatial data to high-dimensional data in the context of content-based image retrieval (CBIR). W.r.t. related work in distributed information retrieval, which is also discussed in this paper, a main characteristic of our approach are more space efficient resource descriptions. This makes them applicable for a wider range of application fields apart from the P2P domain.

1 Introduction

In recent years, there has been a tremendous increase in (personal) web data. Web users maintain blogs, twitter their lives and upload photos and videos to social media sites. Besides storing media items, people tend to share them with friends and interact with each other by collaboratively tagging or commenting on various items. Consequently, heterogeneous online resources which differ in size, media type and update characteristic have to be administered [Thomas and Hawking, 2009]. Hence, effective and efficient retrieval techniques are essential.

Several criteria can be employed for the retrieval of media items (cf. Fig. 1): *a)* textual content *b)* geographic footprints, *c)* timestamps and *d)* (low-level) audio or visual content information. Based on these criteria, text, image, audio and video documents can be indexed and searched.

Peer-to-peer (P2P) scenarios for the administration of media collections are attractive for multiple reasons. Media items can reside on individual devices without a need to store them on remote servers hosted by service providers.

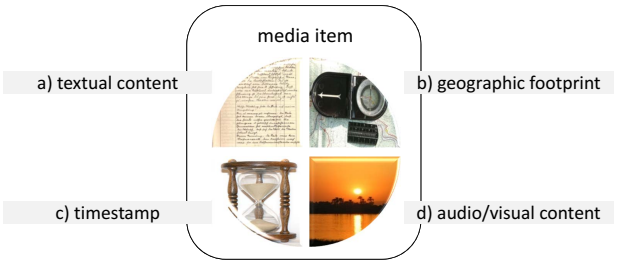

Figure 1: Possible criteria for media retrieval.

Besides reducing dependency from service providers as informational gatekeepers, no expensive infrastructure has to be maintained by applying a scalable P2P protocol such as Rumorama [Müller *et al.*, 2005b]. Crawling, which consumes large amounts of web traffic [Bockting and Hiemstra, 2009], can thus be avoided. Idle computing power in times of inactivity can be used to maintain, analyze and enrich media items.

Our work focuses on space efficient resource description and corresponding selection techniques which allow for efficient and effective query processing. As a proof-of-concept, we design them for the use within Rumorama without limiting their possible application. The resource description and selection techniques can also be applied in the context of traditional distributed information retrieval (IR) (cf. Sect. 2.1) or other variants of P2P IR systems (cf. Sect. 2.2). Furthermore, there is a range of possible application fields apart from P2P IR systems (cf. Sect. 2.3).

Rumorama is a scalable P2P protocol that builds hierarchies of PlanetP-like [Cuenca-Acuna *et al.*, 2003] P2P networks. In Rumorama, every peer sees a portion of the network as a single, small PlanetP network and furthermore maintains connections to other peers that see other small PlanetP networks. To this end, the peer stores a small set of links pointing to neighboring peers in other subnets in order to be able to forward queries beyond the boundaries of its own PlanetP-like subnet. Each peer can choose the size of its PlanetP network according to local processing power and bandwidth capacity. Within its small PlanetP-like subnet, a peer knows resource descriptions of all other peers' data in the same subnet. These descriptions are disseminated by randomized rumor spreading and provide the basis for query routing decisions, i.e. which peers to contact in the local subnet during query processing.

Peers storing media items which are described by the criteria outlined in Fig. 1 can thus be summarized by corresponding resource or peer descriptions (cf. Fig. 2), where each peer is considered as a resource of potentially use-

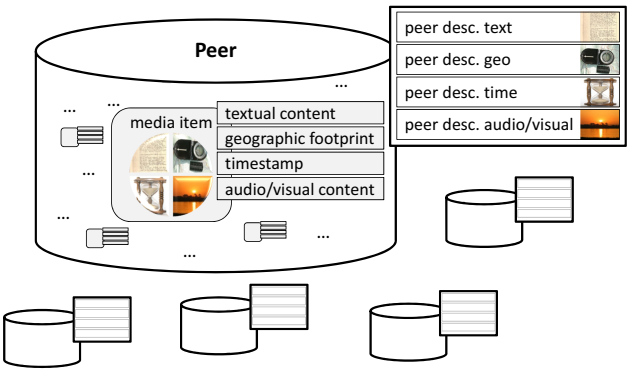

Figure 2: Peer or resource descriptions based on media item features.

ful media items with respect to the given query. These peer or resource descriptions can be envisaged as an aggregation of the features of the media items stored on the respective peer. Resource description and selection for textual data has already been extensively discussed in literature (cf. [Cuenca-Acuna *et al.*, 2003]). Techniques for time and date information are presumably less challenging and might consist of a combination of clustering (cf. [Duda *et al.*, 2000]) and histogram techniques (cf. [Ioannidis, 2003]).

We proposed resource description and selection techniques for geographic data in [Blank and Henrich, 2009; 2010]. Techniques for CBIR were e.g. addressed in [Blank *et al.*, 2007]. In this paper we will apply and adapt ultra fine-grained summaries (UFS, cf. Sect. 3), the most promising technique in the context of geographic data based on binary histograms, for CBIR and analyze its use in more detail. The contribution of this paper is *i)* the application and adaptation of UFS for CBIR, *ii)* a detailed analysis of summary sizes in the case of CBIR, and *iii)* an analysis of time complexity for peer ranking.

Of course, resource ranking based on a single criterion is only a first step. When querying for multiple criteria, e.g. for a sunset image in a certain geographic region, criterion-specific resource rankings can be combined applying a merging algorithm for ranked lists (cf. [Belkin *et al.*, 1995; Ilyas *et al.*, 2008]).

The remainder of this paper is organized as follows. Sect. 2 discusses related work. In Sect. 3 we present an analysis of the application and adaptation of UFS for CBIR. Sect. 4 concludes with an outlook on future work.

2 Related Work

2.1 Traditional Distributed IR Systems

Traditional distributed IR is mainly concerned with text data. Resources are usually described by the set of terms (or a subset of them) which are contained in the documents of a resource and some kind of frequency information per term plus possibly additional statistics [Callan, 2000]. There is plenty of work on resource selection in this context (references are e.g. given in [Thomas and Hawking, 2009; Bockting and Hiemstra, 2009]).

We will now describe approaches which address resource description and selection based on low-level visual content features. Chang et al. [Chang and Zhang, 1997; Chang *et al.*, 1997] propose three approaches. For all of them, a relatively small number of feature vectors from reference images is used (so called templates or icons). In contrast to our approach (cf. Sect. 3), templates are selected from the underlying data collection, i.e. the images that are administered by the resources. For all three approaches, a set of matching templates is computed per image and per query based on a predefined similarity threshold. The first approach uses the number of images which are assigned to a certain template as a ranking criterion. The second approach additionally applies mean and variance information w.r.t. the similarity values between a template and the images which are assigned to it. In experimental studies, both approaches perform worse than a histogram-based approach. Here, a special form of clustering is applied in order to further partition the feature space covered by a template. This approach is e.g. different to our approach w.r.t. the type of clustering, the mechanism for computing the histogram (more parameters needed, no compression, no external collection), and especially the number of applied reference points.

[Berretti *et al.*, 2004] apply a special form of hierarchical clustering to a resource's set of feature vectors in order to ascertain a resource description. A predefined maximum cluster radius is used for determining the centroids which are included in the resource description. Every path in the clustering tree is descended as long as the cluster radius of a node is bigger than the maximum cluster radius. Amongst other information, the centroids of the nodes where the search stops are included in the resource description. By varying the maximum cluster radius, the granularity and size of the resource descriptions can be adapted. When it comes to resource selection, centroids and cluster radii are applied. Compared to our approach, the size of the resource descriptions is expected to be bigger with smaller potential for compression, since centroids are usually represented by d-dimensional real-valued feature vectors (d often between 10^2 and 10^3). We proposed similar approaches in [Alali *et al.*, 2008] where local clustering and Gaussian mixture models (GMMs) are compared. For local clustering d-dimensional cluster centroids are included in the summaries. Mean and variance vectors of dimensionality d plus size information capturing the number of images of a peer which lie in a certain cluster are used as summaries in case of GMMs. GMMs perform better than local clustering in terms of ranking selectivity, but cannot outperform the approaches presented in Sect. 3. In addition, average summary sizes are expected to be bigger since the summarization of a peer with only a single real-valued image feature vector and thus one centroid will consume $d \cdot 4$ bytes for local clustering and $(2d + 1) \cdot 4$ bytes for GMMs, if we assume the usage of 4 bytes per information unit.

[Kim *et al.*, 2002] apply a multi-dimensional selectivity estimation approach based on compressed histograms (cf. [Lee *et al.*, 1999]) for resource description and selection. The d-dimensional feature space is partitioned based on a uniform, multi-dimensional grid. The histogram captures the number of features which lie in a certain bucket of the grid. Since the number of buckets rapidly increases with increasing d, multi-dimensional discrete cosine transform (DCT) is applied in order to reduce histogram sizes. With the help of an adequate sampling strategy, only the most important DCT coefficients are selected for representation. 8 bytes are used per DCT coefficient (4 bytes for the histogram index and 4 bytes for the histogram value). In their experiments in [Kim *et al.*, 2002] $2,000$ till $2,500$ coefficients are used. The entity responsible for resource selection can apply inverse DCT in order to recover the his-

togram with low error rates. For multi-dimensional range queries, the hyper-sphere representing the search region is approximated by multiple hyper-squares. Histogram information is used to determine the selectivity of individual resources w.r.t. the query. The method is further extended in order to support resource selection in heterogeneous settings, where each resource may use its own local similarity measure which may be different from a global similarity measure used by the entity performing resource selection. In [Kim and Chung, 2003], different resource selection strategies are evaluated which do not rely on histogram information and instead use queried feature vectors in order to determine relevant resources by applying different regression models on distributions of global and local similarity value pairs of queried feature vectors.

2.2 P2P IR Systems

P2P IR systems are often classified as being *structured* or *unstructured* overlay networks. As a secondary classification criterion, we introduce the distinction between *data-independent* and *data-dependent* overlays in order to reflect if a peer's content or e.g. query profiles have an effect on overlay generation. This distinction is helpful to pinpoint different characteristics in a more organized way. In the following, we will briefly discuss various approaches.

Unstructured P2P IR Systems

Data-independent: Main protocols in this group are PlanetP [Cuenca-Acuna *et al.*, 2003] and its extension Rumorama [Müller *et al.*, 2005b]. In Rumorama, a peer sees the network as a single, small PlanetP network (called subnet) with connections to other peers that 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 same subnet. Gossiping techniques are used to disseminate the 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. But, Rumorama can be easily extended by a grouping of peers similar to the content-dependent overlays described in the following. Additionally, summaries might be visualized and thus be beneficial 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 (for references cf. [Doulkeridis *et al.*, 2009]) represent aggregated information in an unstructured network 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, they have e.g. been extended to allow for multi-dimensional queries.

Data-dependent: Many semantic overlay networks (SONs) (for references and a detailed description cf. [Doulkeridis *et al.*, 2010]) can be characterized as 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: short links grouping peers with similar content or query profiles into so called "clusters of interest" (COIs) and long links that are established between different COIs. During query execution the query has to be forwarded to the most promising COI(s). In order to form COIs, clustering, classification as well as gossiping techniques can be applied.

Structured P2P IR Systems

Data-independent: Structured P2P IR systems are based on distributed indexing structures with distributed hashtables (DHTs) being 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, indexing 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 CBIR (cf. [Novak *et al.*, 2008; Lupu *et al.*, 2007; Vu *et al.*, 2009]) are usually mapped to one-dimensional or multi-dimensional keys which can be indexed in a classical DHT such as Chord [Stoica *et al.*, 2001] or CAN [Ratnasamy *et al.*, 2001] respectively.

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]. Clustering, classification as well as gossiping techniques are applied in order to establish links to peers with similar content.

Indexing of Multiple Criteria

In structured, data-independent systems, correlations between different criteria (e.g. geographic and image content information) are difficult to exploit when indexing multiple feature types. 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. Within SONs, the simultaneous indexing of multiple criteria 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. Within unstructured, data-independent P2P IR systems, it is possible to apply one summary 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]. Alternatively, the creation of summaries and resource selection algorithms integrating multiple feature types is possible (cf. [Hariharan *et al.*, 2008]).

Hybrid Approaches and Super-Peer Architectures

A main characteristic of unstructured P2P IR systems is that a peer only administers indexing data of media items which belong to its user. Thus, when entering the system or updating media items, full indexing 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, super-peer architectures [Papapetrou *et al.*, 2007] as well as DHT-based indexing of compact data summaries instead of full indexing data has 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 e.g. evaluated an approach where indexing data is stepwisely transferred amongst peers 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 networks. Also within super-peers the convergence of different approaches can be seen. [Doulkeridis *et al.*, 2009] e.g. apply multi-dimensional routing indexes on a super-peer level and additionally group similar super-peers close together in order to allow for better query routing.

In this context, our resource selection techniques are not restricted to data-independent, unstructured P2P IR systems. The summaries can also be used within data-dependent, unstructured P2P IR systems to form COIs and within structured networks e.g. to be indexed in a DHT. In addition, summaries could be used by super-peers for selecting either "normal" peers or other super-peers. Further application fields are also possible as will be described in the following section.

2.3 Possible Application Fields apart from the P2P Context

In addition to P2P IR (cf. Sect. 2.2) our resource summarization and selection techniques can also be used in traditional distributed IR applications (cf. Sect. 2.1). Personal meta-search is a novel application of distributed IR, where all the online resources of a person are queried (e-mail accounts, web pages, image collections, etc.). These resources are typically heterogeneous in size, media type and update frequency possibly requiring selective and space efficient summaries in this context [Thomas and Hawking, 2009].

Our summarization techniques might also be applied within (visual) sensor [Elahi *et al.*, 2009] as well as ad hoc networks [Lupu *et al.*, 2007]. Within 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. [Lupu *et al.*, 2007] present an 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 transfer complete indexing data but only summarized information.

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 a textual query "music beatles" might also be interested in music videos and thus the results of video search or small previews should be integrated in result presentation of classical web search. 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 resource descriptions might also be beneficial in the context of recommender systems and social search e.g. 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 media content.

Another potential application area is automatic theme identification. Automatic theme identification of photo sets e.g. in case of digital print products[1] is concerned with the task of finding suitable background themes for a given set of images. Themes can be travel, wedding, etc. Each theme can be described by a set of photos and modeled as a summary. Afterwards, the theme descriptions and the description of a user's collection can be compared in order to recommend the best matching theme(s).

Resource selection techniques have been successfully applied to blog site search [Elsas *et al.*, 2008]. A blog feed is viewed as a single collection and individual posts are interpreted as documents in order to retrieve similar blogs according to a given information need. A similar approach can be undertaken in passage retrieval such as XML retrieval where different sections, subsections, etc. might be grouped together as a resource (for references cf. [Lalmas, 2009]).

Also expert search [Balog *et al.*, 2009] could presumably be built based on resource description and selection techniques. Here, a user is interested in finding human experts in an enterprise for example. Thus, e.g. all documents a person has (co)authored could be modeled as a resource and finding an expert would result in selecting the most promising resource.

Compact resource descriptions might also be valuable for focused crawling [Ahlers and Boll, 2009]. If a service provides summaries of the image content 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 sets of images in order to extract CBIR features can thus be avoided.

Tree-based index structures are also related to our work (cf. [Samet, 2006]). The decision of choosing the best subtree is similar to the resource selection problem. Summaries in the P2P context correspond to aggregations maintained in the nodes of a tree, e.g. bounding boxes in the case of an R-tree [Guttman, 1984].

3 Resource Summarization and Selection for low-level Visual Content Features

[Müller *et al.*, 2005a] proposed the use of "cluster histograms" for distributed CBIR. In order to compute cluster histograms, a moderate number of reference points is used (e.g. $k = 256$). This set of reference points is known to all peers. Every image feature vector of a peer's local image collection is assigned to the closest reference point. Hereby, a cluster histogram is computed counting how many image feature vectors of a peer's collection are

[1] *http://comminfo.rutgers.edu/conferences/mmchallenge/2010/03/11/cewe-challenge/*, last visit: 8.7.2010

closest to a certain reference point, i.e. cluster centroid c_j ($1 \leq j \leq k$). Reference points are determined by distributed k-Means clustering which imposes some load on the network. During resource selection only histogram information regarding the cluster whose reference point lies closest to the query feature vector is used. Peers with more feature vectors assigned to this cluster are ranked higher than peers with fewer feature vectors assigned to the cluster.

In [Eisenhardt *et al.*, 2006] the performance of resource selection is further improved. A list L of reference points c_j is sorted in ascending order according to the distance of c_j to the query feature vector q. In order to rank peer p_a before p_b or vice versa L is processed from the beginning possibly till the end. The first element of L corresponds to the cluster centroid being closest to q. A peer with more documents in this cluster is ranked higher than a peer with fewer documents in the very cluster. If two peers p_a and p_b administer the same amount of images in the analyzed cluster and the end of the list has not yet been reached, the next element out of L is chosen and based on the number of documents within the current cluster it is again tried to rank p_a before p_b or vice versa. As a second modification, distributed clustering is replaced in [Eisenhardt *et al.*, 2006] by a random selection of reference points. Overall ranking selectivity is slightly affected, but there is no longer any network load imposed due to distributed clustering.

We have extended the work from [Müller *et al.*, 2005a; Eisenhardt *et al.*, 2006] in several directions (cf. [Blank *et al.*, 2007]). First, within highly fine-grained summaries (HFS_k with k indicating the number of reference points used) we increased the number of reference points for computing the cluster histogram, e.g. from $k = 256$ to $k = 8,192$ or even more. By doing so, the feature space is partitioned in a more fine-grained way offering improved ranking selectivity. Since a higher number of reference points would lead to less space efficient summaries, we apply compression techniques. Thus, we can achieve better ranking selectivity with more space efficient resource descriptions compared to the approaches in [Müller *et al.*, 2005a; Eisenhardt *et al.*, 2006]. The average size of a peer's summary information is approx. 110 bytes for $k = 16,384$, which is clearly less compared to other approaches in distributed CBIR (cf. Sect. 2.1) if they were applied to our scenario directly. Second, within our approach reference points are selected from an external source and transferred to peers together with updates of the P2P software. This leads to a decrease in overall network load and makes distributed selection mechanisms obsolete. Ranking selectivity is slightly affected by this change as will be shown in Sect. 3.3.

3.1 Experimental Setup

In the experiments we use a 166-dimensional uniformly quantized color histogram[2] based on the HSV color space with 18 hues, 3 saturations and 3 values, plus 4 levels of gray. Image feature vectors are compared using Euclidean distance. We perform 20 runs where we change the reference points used. During a run we perform 100 queries where we randomly select a query image from the underlying collection. The set of queries stays constant over all runs. By analyzing the number of peers which are contacted on average in order to retrieve the 20 closest feature vectors w.r.t. a given query feature vector we assume that the most similar images are the ones the user is interested in. We crawled a collection of $233,827$ Flickr images. They are assigned to peers based on the Flickr user ID in order to reflect a realistic scenario. Hence, we assume that every Flickr user operates a peer of its own. The images are mapped to $10,601$ peers/users which are used in our simulation. Fig. 3 shows the distribution of peer sizes, i.e. the number of images which are maintained per peer. The general characteristic is typical for P2P file sharing applications with few peers managing large amounts of the images and many peers administering only few images [Saroiu *et al.*, 2002].

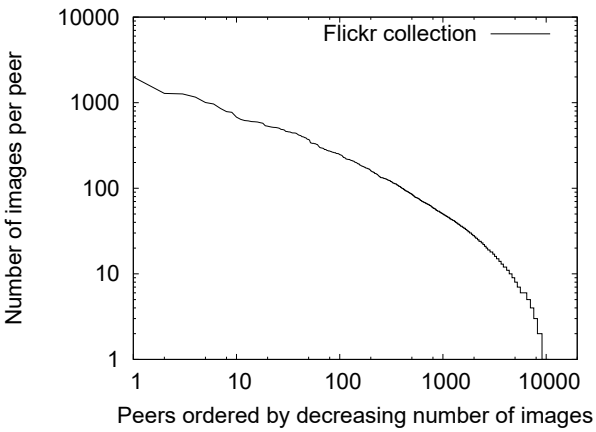

Figure 3: Distribution of peer sizes.

3.2 Using UFS for CBIR

When summarizing geographic footprints, binary histograms (so called UFS_k: ultra fine-grained summaries) outperformed HFS_k (cf. [Blank and Henrich, 2010]). 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 feature vectors. Hence, we obtain a bit vector of size k. Of course, there is some loss of information when switching from HFS to UFS with k 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 ranking selectivity compared to HFS. In the following we will thus evaluate the use of UFS in the context of high-dimensional feature vectors for CBIR.

3.3 Analysis of Ranking Selectivity

Reference points for summary creation and peer ranking are chosen from the underlying collection (UFS/HFS) or a second collection of $45,931$ Flickr images (UFSe/HFSe with "e" indicating the use of an external collection for the reference points). It is important to note that both collections are disjoint w.r.t. the unique Flickr image and user IDs, but there is some minor natural overlap amongst collections w.r.t. image content; 24 of the $233,827$ images also appear in the external collection, because some images are uploaded by multiple users independently on Flickr.

Fig. 4 shows the number of peers which are contacted on average in order to retrieve the 20 closest feature vectors w.r.t. a given query feature vector. Ranking selectivity increases degressively with increasing k. There is a gap in ranking selectivity when choosing the reference points

[2] *http://www.gnu.org/software/gift/*, last visit: 5.7.2010

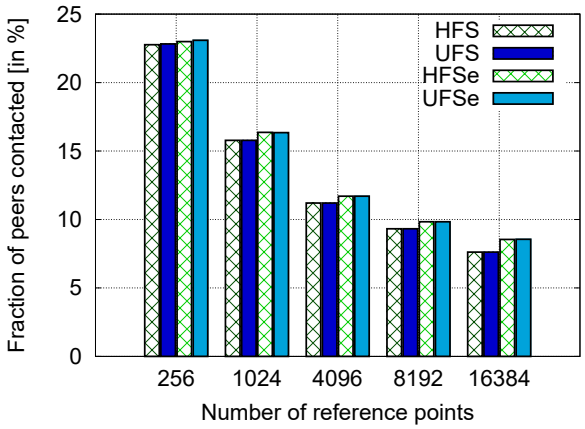

Figure 4: Fraction of peers contacted to retrieve top-20 image feature vectors (ranking selectivity).

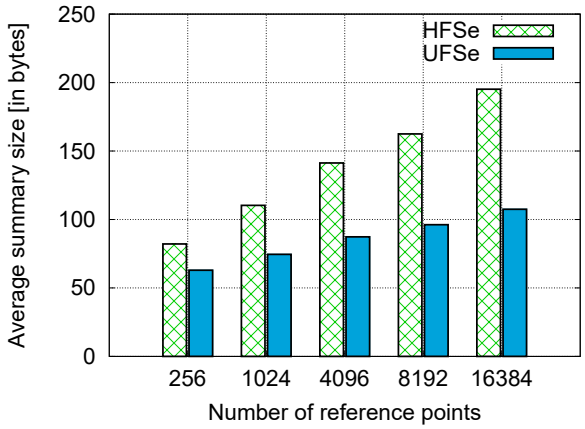

Figure 5: Avg. summary sizes (zipped).

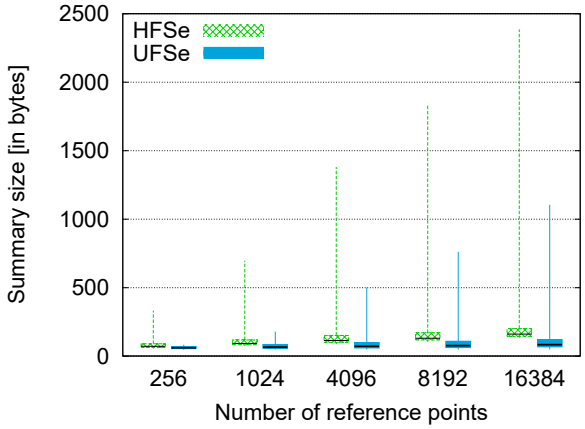

Figure 6: Box plot of summary sizes.

from an external collection (HFSe compared to HFS and UFSe compared to UFS). The gap increases with increasing k. For UFS/HFS, with increasing k also the probability of choosing a centroid which is used also as a query feature vector increases. Such situations might lead to improved ranking selectivity, since queries are randomly chosen from the underlying data collection. The evaluation of other sources of feature vectors as queries will be part of future work. Fig. 4 additionally shows slightly improved ranking selectivity for HFS(e)$_{256}$ compared to UFS(e)$_{256}$ respectively[3], which is due to the use of non-binary histogram information during peer ranking within HFS(e). In general, this gap more and more diminishes when increasing the number of centroids used since HFS(e) histograms more and more pass into binary histograms. Already for HFS$_{1,024}$ compared to UFS$_{1,024}$ and HFSe$_{1,024}$ compared to UFSe$_{1,024}$ there is no noticeable difference in ranking selectivity at all.

3.4 Analysis of Summary Sizes

The size of the resource descriptions after zipping is analyzed in Fig. 5 and Fig. 6. Fig. 5 shows average summary sizes s_{avg} when using UFSe instead of HFSe. The plot for UFS and HFS shows similar characteristics. In addition, Fig. 6 visualizes the different quartiles and minimum/maximum values of the summary sizes in a box plot. It shows that the median in case of HFSe is bigger compared to UFSe. Interquartile ranges of HFSe and UFSe become more and more similar when increasing k although the overall range of HFSe summary sizes is greater than the range of summary sizes in case of UFSe. All distributions of summary sizes are positively skew indicating many peers with small summary sizes and few peers with big summary sizes. Thus, the distribution of peer sizes (cf. Fig. 3) is reflected in the distribution of summary sizes (cf. Fig. 6).

One might think of a hybrid peer ranking scheme e.g. using HFSe for the smaller peers (i.e. peers with few documents) and UFSe for the bigger peers (i.e. peers with many documents) in order to reduce network load imposed by rumor spreading. The cost of one round of rumor spreading can be estimated by $s_{avg} \cdot n \cdot (n-1)$ with n being the number of peers in a PlanetP-like network. Hence, the cost is proportional to s_{avg}. Since an increase in ranking selectivity can be perceived only for small values of k when switching from HFS(e) to UFS(e) respectively (cf. Fig. 4), the UFS(e) alternative can be safely chosen for all peers in the network in case of big values of k. If there are bigger differences in ranking selectivity amongst competing resource description and peer ranking schemes, the cost for query processing has to be additionally taken into account. A more detailed analysis can be found in [Blank and Henrich, 2010].

3.5 Analysis of Time Complexity

In general, it is important that peer ranking can be done within a reasonable amount of time. As described above, ranking peers mainly means sorting k-dimensional numbers where the importance of the single dimensions is defined by the list L which contains the reference points sorted according to their distance to the query feature vector. In a first run the peers are sorted w.r.t. the dimension representing the closest reference point. Of course, this sorting can be done in $\mathcal{O}(n \cdot \log n)$ where n stands for the number of peers in the considered PlanetP network. In a worst case scenario all peers would be identical in the number of media items maintained in each of the k clusters ending up in a complexity of $\mathcal{O}(k \cdot n \cdot \log n)$. Thus, the worst case complexity for calculating a peer ranking depends on k which of course is disadvantageous for HFS(e)/UFS(e) with high values of k.

[3] HFS(e)$_k$ is used as an abbreviation for "HFS$_k$ and HFSe$_k$". The same notation is also adopted for UFS throughout the paper. In a similar way, HFS/UFS also abbreviates "HFS and UFS".

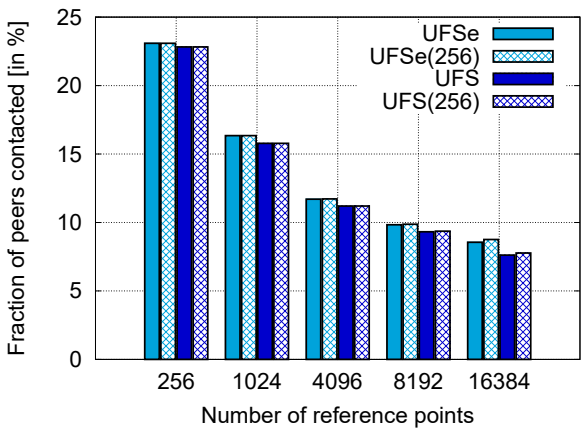

Figure 7: Fraction of peers contacted for retrieving top-20 images with UFS(e)(256) vs. UFS(e) (ranking selectivity).

In order to test whether this worst case scenario has practical implications, we compared the original approach considering the clusters to the end, if necessary, with a modified variant using at most the 256 clusters closest to the query. If no decision is possible after considering the histogram values for these 256 clusters, a random choice is made. In the following UFS(e)$_k$(256) will denote the modified approach considering k centroids for summary creation while at most only applying the 256 closest centroids w.r.t. the query feature vector for peer ranking. Results can be seen in Fig. 7. There is no noticeable difference in ranking selectivity for UFS(e)$_k$ compared to UFS(e)$_k$(256) respectively for summaries with up to $k = 8,192$ centroids. UFS(e)$_{16,384}$ performs slightly better than UFS(e)$_{16,384}$(256). When increasing k, the feature space is partitioned in a more fine grained way. If only 256 centroids are used for peer ranking, the fraction of unused centroids which potentially contain relevant information increases, e.g. in case of UFS(e)$_{16,384}$(256), $1 - \frac{256}{16,384} = 98.4\%$ of summary information is discarded during peer ranking.

The results in Fig. 7 demonstrate two things. First, obviously very few of the k histogram bins are usually considered for peer ranking. Otherwise the differences between UFS(e) and UFS(e)(256) would have been higher. Second, programmers anxious about worst case bounds can stop processing after considering a certain number of histogram bins and thus avoid the worst case of $\mathcal{O}(k \cdot n \cdot \log n)$.

4 Conclusion & Outlook

In this paper we have applied and adapted binary histograms which were originally designed for the summarization of low-dimensional spatial data for resource description and selection based on high-dimensional CBIR features. Compared to earlier work, summaries can be zipped more efficiently. A huge number of reference points (e.g. $16,384$) is applied in order to generate resource descriptions. We have shown that it is possible to use only a small fraction of reference points (e.g. 256) during peer ranking in order to speed-up query processing with only a marginal decrease in ranking selectivity.

In future work we will try to find an adequate stopping criterion which indicates when it is no longer beneficial to contact further peers. This might be woven with a technique that adaptively determines the number of centroids used for peer ranking. Additionally we will apply our resource descriptions in order to summarize local image features such as SIFT.

References

[Ahlers and Boll, 2009] Dirk Ahlers and Susanne Boll. Adaptive geospatially focused crawling. In *Proc. of the 18th ACM Conf. on Information and Knowledge Management*, pages 445–454, Hong Kong, China, 2009.

[Allali *et al.*, 2008] Soufyane Allali, Daniel Blank, Wolfgang Müller, and Andreas Henrich. Image data source selection using Gaussian mixture models. In *Adaptive Multimedial Retrieval: Retrieval, User, and Semantics: 5th Intl. Workshop, AMR 2007, Paris, France, 2007, Revised Selected Papers*, pages 170–181, Berlin, Heidelberg, 2008. Springer LNCS 4918.

[Arguello *et al.*, 2009] Jaime Arguello, Fernando Diaz, Jamie Callan, and Jean-Francois Crespo. Sources of evidence for vertical selection. In *Proc. of the 32nd Intl. ACM SIGIR Conf. on research and development in Information Retrieval*, pages 315–322, Boston, MA, USA, 2009.

[Balog *et al.*, 2009] Krisztian Balog, Ian Soboroff, Paul Thomas, Nick Craswell, Arjen de Vries, and Peter Bailey. Overview of the TREC 2008 enterprise track. In *The Seventeenth Text Retrieval Conf. Proceedings (TREC 2008)*. NIST, 2009. Special Publication.

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

[Bender *et al.*, 2005] Matthias Bender, Sebastian Michel, Gerhard Weikum, and Christian Zimmer. The minerva project: Database selection in the context of P2P search. In *Datenbanksysteme in Business, Technologie und Web, 11. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS)*, pages 125–144, Karlsruhe, Germany, 2005.

[Berretti *et al.*, 2004] Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. Merging results for distributed content based image retrieval. *Multimedia Tools Appl.*, 24(3):215–232, 2004.

[Blank and Henrich, 2009] Daniel 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.

[Blank and Henrich, 2010] Daniel Blank and Andreas Henrich. Description and selection of media archives for geographic nearest neighbor queries in P2P networks. In *Proc. of IAPMA2010: Information Access for Personal Media Archives Workshop*, pages 22–29, `http://doras.dcu.ie/15373/` (25.5.2010), Milton Keynes, UK, 2010.

[Blank *et al.*, 2007] Daniel Blank, Soufyane El Allali, Wolfgang Müller, and Andreas 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.

[Bockting and Hiemstra, 2009] Sander Bockting and Djoerd Hiemstra. Collection Selection with Highly Discriminative Keys. In *Proc. of the 7th Workshop on Large-Scale Distributed Systems for Information Retrieval*, `http://lsdsir09.isti.cnr.it/lsdsir09-1.pdf` (26.04.2010), Boston, MA, USA, 2009.

[Callan, 2000] Jamie Callan. Distributed information retrieval. In W. B. Croft, editor, *Advances in Information Retrieval*, pages 127–150. Kluwer Academic Publishers, 2000.

[Chang and Zhang, 1997] Wendy Chang and Aidong Zhang. Metadata for distributed visual database access. In *2nd IEEE Metadata Conf.*, pages 1–11, Silver Spring, MD, USA, 1997.

[Chang *et al.*, 1997] Wendy Chang, Gholamhosein Sheikholeslami, Aidong Zhang, and Tanveer F. Syeda-Mahmood. Efficient resource selection in distributed visual information systems. In *Proc. of the 5th ACM Intl. Conf. on Multimedia*, pages 203–213, Seattle, Washington, USA, 1997.

[Cuenca-Acuna *et al.*, 2003] Francisco Cuenca-Acuna, Christopher Peery, Richard P. Martin, and Thu D. Nguyen. PlanetP: Using gossiping to build content addressable peer-to-peer information sharing communities. In *IEEE Intl. Symp. on High Performance Distributed Computing*, pages 236–246, Seattle, WA, USA, 2003.

[Doulkeridis *et al.*, 2009] Christos Doulkeridis, Akrivi Vlachou, Kjetil Nørvåg, Yannis Kotidis, and Michalis Vazirgiannis. Multidimensional routing indices for efficient distributed query processing. In *Proc. of the 18th ACM Conf. on Information and Knowledge Management*, pages 1489–1492, Hong Kong, China, 2009.

[Doulkeridis *et al.*, 2010] Christos Doulkeridis, Akrivi Vlachou, Kjetil Nørvåg, and Michalis Vazirgiannis. *Handbook of Peer-to-Peer Networking*. Part 4: Distributed Semantic Overlay Networks. Springer Science+Business Media, 1st edition, 2010.

[Duda *et al.*, 2000] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley-Interscience, 2nd edition, Nov. 2000.

[Eisenhardt *et al.*, 2006] Martin Eisenhardt, Wolfgang Müller, Andreas Henrich, Daniel Blank, and Soufyane El Allali. Clustering-based source selection for efficient image retrieval in peer-to-peer networks. In *8th IEEE Intl. Symp. on Multimedia*, pages 823–830, San Diego, CA, USA, 2006.

[Eisenhardt *et al.*, 2008] Martin Eisenhardt, Wolfgang Müller, Daniel Blank, Soufyane El Allali, and Andreas Henrich. Clustering-based, load balanced source selection for CBIR in P2P networks. *Intl. Journal of Semantic Computing (IJSC)*, 2(2):235–252, 2008.

[Elahi *et al.*, 2009] B. Maryam. Elahi, Kay Römer, Benedikt Ostermaier, Michael Fahrmair, and Wolfgang Kellerer. Sensor ranking: A primitive for efficient content-based sensor search. In *Intl. Conf. on Information Processing in Sensor Networks*, pages 217–228, Washington, DC, USA, 2009. IEEE.

[Elsas *et al.*, 2008] Jonathan L. Elsas, Jaime Arguello, Jamie Callan, and Jaime G. Carbonell. Retrieval and feedback models for blog feed search. In *Proc. of 31st Intl. ACM SIGIR Conf. on research and development in Information Retrieval*, pages 347–354, Singapore, 2008.

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

[Guy *et al.*, 2010] Ido Guy, Michal Jacovi, Adam Perer, Inbal Ronen, and Erel Uziel. Same places, same things, same people?: mining user similarity on social media. In *CSCW '10: Proc. of ACM Conf. on Computer Supported Cooperative Work*, pages 41–50, Savannah, Georgia, USA, 2010.

[Hariharan *et al.*, 2008] Ramaswamy Hariharan, Bijit Hore, and Sharad Mehrotra. Discovering GIS sources on the web using summaries. In *Proc. of 8th ACM/IEEE joint Conf. on digital libraries*, pages 94–103, Pittsburgh, PA, USA, 2008. ACM.

[Ilyas *et al.*, 2008] Ihab F. Ilyas, George Beskales, and Mohamed A. Soliman. A survey of top-k query processing techniques in relational database systems. *ACM Comput. Surv.*, 40(4):1–58, 2008.

[Ioannidis, 2003] Yannis Ioannidis. The history of histograms (abridged). In *Proc. of 29th Intl. Conf. on Very Large Data Bases*, pages 19–30, Berlin, Germany, 2003.

[Kim and Chung, 2003] Deok-Hwan Kim and Chin-Wan Chung. Collection fusion using Bayesian estimation of a linear regression model in image databases on the web. *Inf. Processing and Management*, 39:267–285, 2003.

[Kim *et al.*, 2002] Deok-Hwan Kim, Seok-Lyong Lee, and Chin-Wan Chung. Heterogeneous image database selection on the web. *The Journal of Systems and Software*, 64:131–149, 2002.

[Lalmas, 2009] Mounia Lalmas. *XML retrieval*. Synthesis Lectures on Information Concepts, Retrieval and Services. Morgan & Claypool Publishers, 2009.

[Lee *et al.*, 1999] Ju-Hong Lee, Deok-Hwan Kim, and Chin-Wan Chung. Multi-dimensional selectivity estimation using compressed histogram information. *SIGMOD Record*, 28(2):205–214, 1999.

[Lupu *et al.*, 2007] Mihai Lupu, Jianzhong Li, Beng Chin Ooi, and Shengfei Shi. Clustering wavelets to speed-up data dissemination in structured P2P manets. In *Intl. Conf. on Data Engineering*, pages 386–395, Istanbul, Turkey, 2007. IEEE.

[Müller *et al.*, 2005a] W. Müller, M. Eisenhardt, and A. Henrich. Fast retrieval of high-dimensional feature vectors in P2P networks using compact peer data summaries. *Multimedia Systems*, 10(6):464–474, 2005.

[Müller *et al.*, 2005b] Wolfgang Müller, Martin Eisenhardt, and Andreas Henrich. Scalable summary based retrieval in P2P networks. In *Proc. of the 14th ACM Conf. on Information and Knowledge Management*, pages 586–593, Bremen, Germany, 2005.

[Novak *et al.*, 2008] David Novak, Michal Batko, and Pavel Zezula. Web-scale system for image similarity search: When the dreams are coming true. In *Intl. Workshop on Content-Based Multimedia Indexing*, pages 446–453, London, UK, 2008. IEEE.

[Papapetrou *et al.*, 2007] Odysseas Papapetrou, Wolf Siberski, Wolf-Tilo Balke, and Wolfgang Nejdl. DHTs over peer clusters for distributed information retrieval. In *21st Intl. Conf. on Advanced Information Networking and Applications*, pages 84–93, Niagara Falls, Canada, 2007. IEEE.

[Ratnasamy *et al.*, 2001] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Schenker. A scalable content-addressable network. In *ACM SIGCOMM*, pages 161–172, San Diego, CA, USA, 2001.

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

[Saroiu *et al.*, 2002] Stefan Saroiu, P. Krishna Gummadi, and Steven D. Gribble. A measurement study of peer-to-peer file sharing systems. In *ACM/SPIE Multimedia Computing and Networking*, pages 156–170, San Jose, CA, USA, 2002.

[Stoica *et al.*, 2001] Ion Stoica, Robert Morris, David Karger, Frans Kaashoek, and Hari Balakrishnan. Chord: A scalable peer-to-peer lookup service for Internet applications. In *SIGCOMM'01*, pages 149–160, San Diego, CA, USA, 2001.

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

[Vu *et al.*, 2009] Quang Hieu Vu, Mihai Lupu, and Sai Wu. Simpson: Efficient similarity search in metric spaces over P2P structured overlay networks. In *Proc. of 15th Intl. Euro-Par Conf. on Parallel Processing*, pages 498–510, Delft, The Netherlands, 2009. Springer.