Diversity in Image Retrieval using Randomization and Learned Metrics

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science (by Research)
in
Computer Science and Engneering

by

P Vidyadhar Rao
201207718
vidyadhar.rao@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
OCTOBER 2015

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Diversity in Image Retrieval using Randomization and Learned Metrics" by P Vidyadhar Rao, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Prof. C.V Jawahar

To

Family and Friends, near and far.

Abstract

Providing useful information to user requested queries is the most investigated problem in the multimedia retrieval community. The problem of information retrieval usually has many possible solutions, due to uncertainities in the user‘s information need and ambiguities in query specification. Some mechanism is required to evaluate the options and select a solution. This is quite a challenging task. In the recent years, the focus is gradually shifting towards *relevance* and *diversity* of retrieved information, which together improve the usefulness of retrieval system as perceived by users. Intuitively it is desirable to design a retrieval system with three requirements: a) Accurate retrieval i.e., the method should have high precision, b) Diverse retrieval, i.e., the obtained results should be diverse, c) Efficient retrieval, i.e., response time should be small. While considerable effort has been expended to develop algorithms which incorporate both relevance and diversity in the retrieval process, relatively less attention has been given to design efficient algorithms.

The main contribution of this thesis lies in developing efficient algorithms for the diverse retrieval problem. We show that the diverse retrieval problem can be mathematically defined as an integer convex optimization problem, and hence finding the optimal solution is NP-Hard. The existing approximate and greedy algorithms that try to find solution to this problem suffer from two drawbacks: a) Running time of the algorithms is very high as it is required to recover several exact nearest neighbors. b) Computations may require an unreasonably large amount of memory overhead for large datasets. In this work, we propose a simple approach to overcome the above issues based on two ideas: 1) Randomization and 2) Learned Metrics

In the first case, the method is based on locality senstive hashing and tries to address all of the above requirements simultaneously. We show that the effectiveness of our method depends on *randomization* in the design of the hash functions. Further, we derive a theoretically sound result to support the intuitiveness and reliability of using hash functions (via randomization) in the retrieval process to improve diversity. We modify the standard hash functions to take into account the distribution of the data for better performance. We also formulate the diverse multi-label prediction (of images and web pages) in this setting and demonstrate the scalability and diversity in the solution.

To validate our proposal and to gain more insights into existing methods, we empirically compare our method against common as well as several diversity-based retrieval methods. We demonstrate effectiveness of our approach in different large-scale retrieval tasks via Image Category Retrieval, Multi-label Classification and Image Tagging.

Our findings show that the proposed hash functions in combination with the existing diversity-based methods significantly outperforms standard methods without using hash functions. Our method allows to achieve a trade-off between accuracy and diversity using easy to tune parameters. We examine evaluation measures for diversity in several retrieval scenarios and introduce a new notion to simultaneously evaluate a method's performance for both the precision and diversity measures. Our proposal does not harm, but instead increases the reliability of the measures in terms of accuracy and diversity while ensuring $100x$-speed-up over the existing diverse retrieval approaches.

In the second case, the method is based on *learning distance metrics*. We show that effectivenesss of our method depends on the learned distance metrics that suits the user‘s perception. In the case of instance based image retrieval methods, relevance and diversity are relative to viewpoint of the camera, time of day, and camera zoom. We argue that the low-level image features fail to capture diversity with respect to high-level human semantics. We use the high-level semantic information to learn metrics and re-fashion the visual feature space to appreciate diversity better. Our proposal is the best strategy from a learning perspective and we empirically demonstrate, when compared to original feature space, that the learned metrics provide better diversity in the retrieval.

In conclusion, in this thesis we discussed two fundamental ideas for retrieving diverse set of results. From the algorithmic and statistical perspective, the proposed method intuitively uses ”randomness as resource” to improve diversity in retrieval while ensuring sub-linear retrieval time. From the visual perspectives, the proposed method utilizes user level semantics to ”learn metrics” for improving diversity in instance based image retrieval. We believe that the ideas presented in this thesis are not limited to image retrieval and therefore, its applicability to different definitions of diversity, knowledge source combination, interactive retrieval systems, and so forth are possible.

Contents

List of Figures

List of Tables

Chapter 1

Diversity in Image Retrieval

> *The world has come a long way since the days of the printing press. Information is no longer a scarce commodity; we have more of it than we know what to do with. But relatively little of it is useful. We perceive it selectively, subjectively, and without much self-regard for the distortions that this causes. We think we want information when we really want knowledge. The signal is the truth. The noise is what distracts us from the truth.* - The Signal and the Noise: Why So Many Predictions Fail but Some Don't [88]

Information has become more and more available on the internet, especially after the advent of several social and community-based services like Wikpedia, Personal blogs, Facebook, Twitter, Instagram, Pinterest etc. Text, images, audio, and videos are just a few of the many media types that are vast quantities of information created and distributed among the users. Users constantly engage in the *information-seeking* activity to explore for new information published on these platforms. However, the available information may contain highly similar or nearly duplicate content. This is because, the same information (for example, popular news) is often created or shared by different users or media outlets. That is, users expect to identify a fairly small number of documents (textual, visual or other), in response to a user's description of their request (e.g., query). Therefore, it is of utmost importance to identify the necessary/useful information from the vast amount of information content.

In the information retrieval context, a typical search involves extracting appropriate features from the query and then perform matching to the instances in the database to find similar results that are relevant to the user query. Increasingly, this technique has drawn more and more attention from the extant web search engines (e.g. Google, Yahoo!, Bing, Facebook graph search, twitter hash tags and so on). The nature of retrieval task in each case is different and also depends on the goals and intentions of the target users. More often than not the user is only vaguely aware of his/her intent.

Most of the retrieval systems are purely similarity based and produce results based on some kind of scalar metric. This, though true for any general retrieval application, manifests itself boldly in image and video retrieval systems. In the context of image retrieval, there are several ways the query can be specified to the retrieval system. The two broad categories of solutions that are of interest are: a)

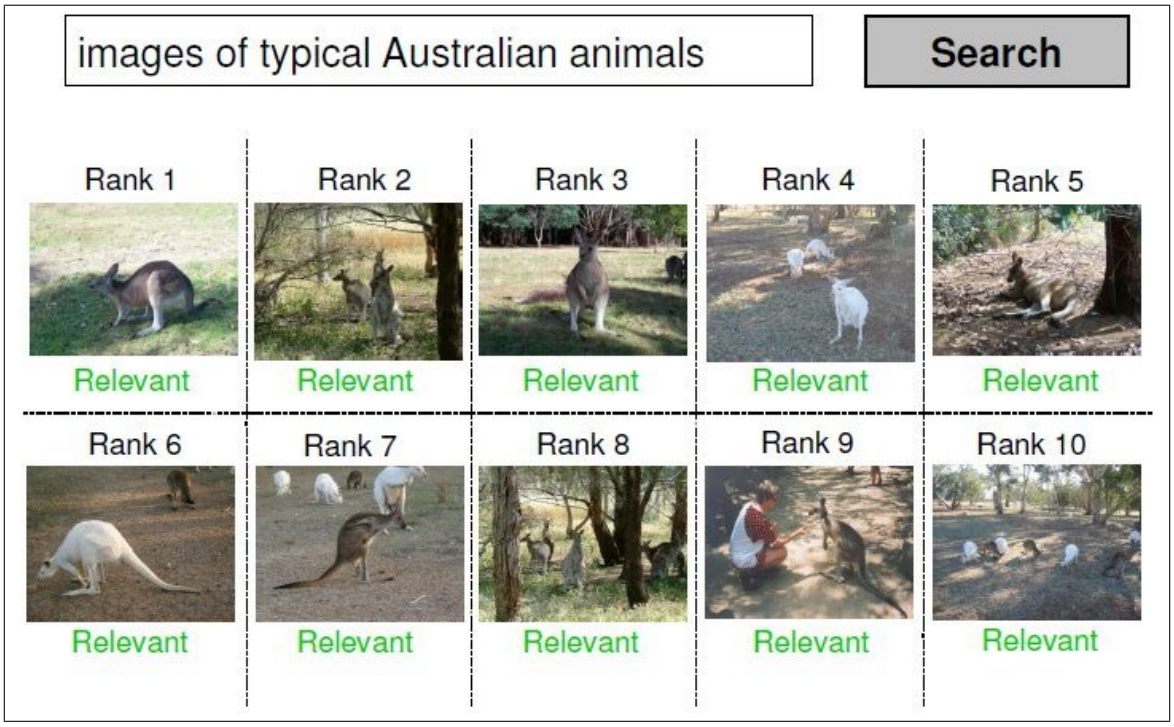


Figure 1.1: Perfect retrieval result with many near duplicates in top ranked images for the query "Australian animals". Source: [4]

Instance based retrieval and b) Category based retrieval. Methods that retrieve the same instance (e.g., images of the same object taken in a different imaging setting) and methods that retrieve images from the same category (e.g., retrieve all flowers or retrieve kitchen images).

Instance based methods are often designed to retrieve a candidate set of images, while the category retrieval methods need more powerful classification than a simple matching. In the case of category retrieval, the problem is closely related to that of content-based image retrieval (CBIR), where the goal is to return better image search results rather than training a classifier for image recognition. For instance, when a user is looking for *Australian animals*, the system would tend to return all the images of *kangaroos*. The user might intend to search for different animals, so it would be more useful to show a variety of animals in the search result. This issue is illustrated in the Figure 1.1 and Figure 1.2. For more details on the why search results need to be diverse can be found in the works [4, 81].

Therefore, ambiguity in user's query (see Figure 1.3) and its inexact interpretation (see Figure 1.4) might often lead to poor user satisfaction, especially when the search results include redundantly similar data. A part from the relevance of the results with regard to query, interactive response within a few seconds is among the foremost criteria for judging the usefulness of an information retrieval system. Moreover, each user's intent is different of the others, and it is better to show both *relevant and diverse* sets of results to maximize the reliablity of the retrieval system. Even then, the system is expected to retrieve results which match the user's intent and that too in interactive time. It is not only a key factor

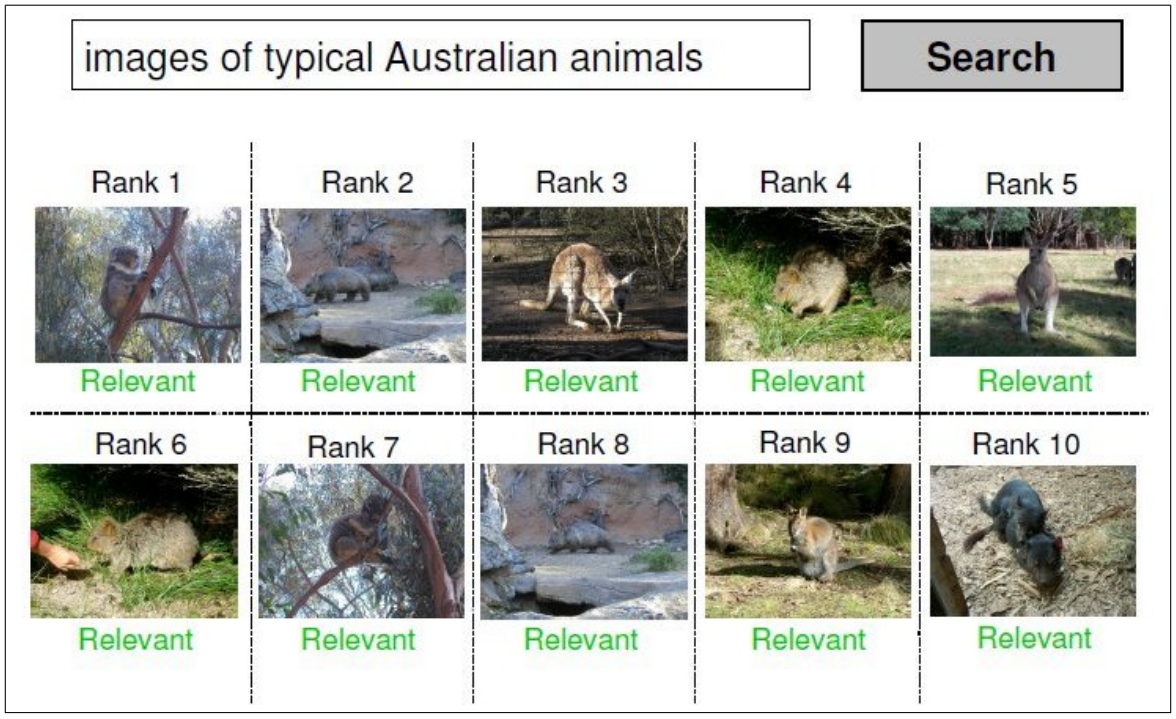


Figure 1.2: Another perfect retrieval result with different animals in Australia for the query "Australian animals". Source: [4]

to address the uncertainity and ambiguity in an information need, but also an effective and efficient way to cover different aspects of the information need [75].

In this thesis, we primarily focus on the retrieval task formulated as the problem of finding nearest neighbors to the user query. This will conceptually simplify the task of analyzing complex modeling problems, thus making it easier to reason about higher level goals and properties of the retrieval system.

1.1 Introduction

Nearest neighbor (NN) retrieval is a critical sub-routine for machine learning, databases, signal processing, and a variety of other disciplines. Basically, we have a database of points, an input query, and the goal is to return the nearest point(s) to the query using some similarity metric. As a naïve linear scan of the database is infeasible in practice, most of the research for NN retrieval has focused on making the retrieval efficient with either novel index structures [16, 107] or by approximating the distance computations [6, 43]. That is, the goal of these methods is: a) accurate NN retrieval, b) fast retrieval.

However, NN retrieval methods [28, 51] are expected to meet one more criteria: diversity of retrieved data points. That is, it is typically desirable to find data-points that are diverse while maintaining high accuracy levels. Our work began by recognizing at the existing diverse retrieval methods that can be broadly categorized into the following two approaches: (a) *Backward selection*: retrieve all the rele-

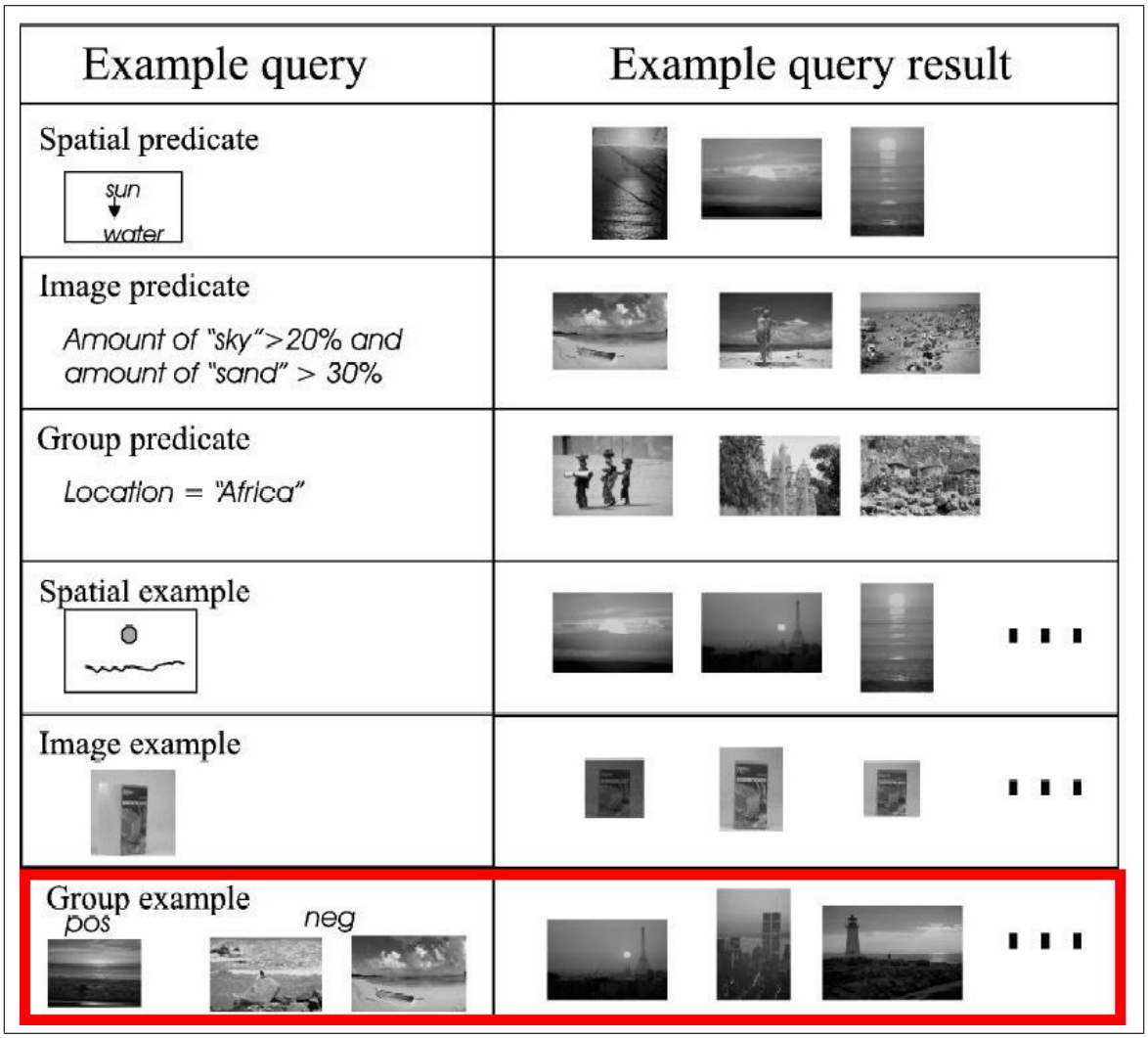


Figure 1.3: Several ways of requesting a query in the context of image retrieval: Users intent is more complex. (Apologies for missing the source to this image)

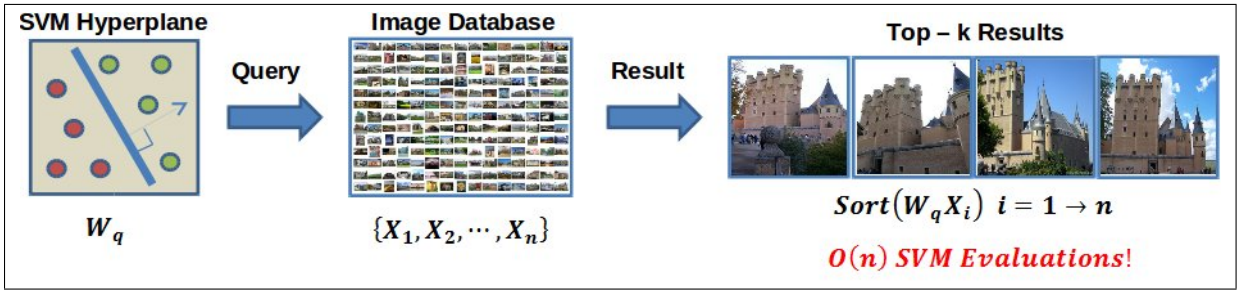


Figure 1.4: Category based image retrieval framework using SVM hyperplane queries: First a classifier is trained using SVM on a set of annotated images for category. Second, the unannotated images are ranked based on the classifier scores. Such classifiers aretoo rigid to interpret the user‘s intent.

vant ones and then find a subset of points with high diversity, (b) *Forward selection*: retrieve points sequentially by optimizing the relevance and diversity scores with a greedy algorithm [13, 26, 39, 110]. However, both these approaches are computationally very expensive than the NN, rendering them infeasible for large scale retrieval applications.

The need for diversity is not limited to retrieval and there has been significant research in many applications [21, 48, 68]. Several research areas and applications use the notion of diversity in a variety of ways. For example, in active learning [21], a diversity measure based on Shannon's entropy is used. Probabilistic models like determinantal point processes [33, 52] collect human diversity judgements using Amazon's Mechanical Turk. Structured SVM based framework [109] measures diversity using subtopic coverage on manually labelled data. In our work, the definition of what constitutes diversity varies across each task and is clearly described. Thus, the evaluation measures used to assess the performance of different methods are also different.

In practise, there is no evaluation metric that seems to be universally accepted as the best for measuring the performance of algorithms that aim to obtain diverse rankings, perhaps in part because, there is a wide diversity in what diversity means [75]. Evaluating these algorithms requires effectiveness measures that appropriately reward diversity in the result list [35, 36]. Diversity also depends on a system's performance at basic ad hoc retrieval i.e., how many points are relevant to any reasonable intent, especially at the top of the ranked list. Therefore, similar to precision and recall, there is a need to balance between accuracy and diversity in the retrieval. In this thesis, we keep a balance between accuracy and diversity by maximizing the harmonic mean of these two criteria. We believe that this performance measure is suitable for several applications and helps us empirically compare different methods.

To this end, we propose a simple retrieval schemes that addresses all of the above mentioned requirements, i.e., a) accuracy, b) retrieval time, c) diversity. The proposed methods are based on two ideas: 1) Randomization and 2) Learned Metrics.

In the first case, the algorithm is based on the following simple observation: in most of the cases, one needs to trade-off accuracy for diversity. That is, rather than finding the nearest neighbor, we would need to select a point which is a bit farther from the given query but is *dissimilar* to the other retrieved points. Hence, we would need to find *approximate nearest neighbors* while ensuring that the retrieved points are diverse. That is, while earlier approaches considered approximate retrieval to be acceptable only for the sake of efficiency, we argue that one can further exploit approximate retrieval to provide impressive trade-offs between accuracy and diversity. We demonstrate that a locality sensitive hashing based randomized algorithm guarantees retrieval in sub-linear time and with superior diversity.

In the second case, the algorithm is based on the following observation: since, relevance and diversity are based on distance metrics, it is crucial to find appropriate metrics when user's view of the data are different. In the context of instance based image retrieval, we define *diversity* as variation of physical properties among most relevant retrieved results for a query image. To achieve this, we *learn* distance metrics that appreciates diversity in images with respect to geometric and illumination properties like viewpoint, time of day and camera zoom.

1.2 Contributions and Roadmap

The main contributions of this thesis can be summarized as follows:

- A systematic investigation of the need for diverse retrieval algorithms from the view of image retrieval is introduced. We formulate the diverse retrieval problem as an integer optimization problem and show that existing methods are computationally expensive, even for the special cases. We propose methods based on two key ideas: *Randomization* and *Learned Metrics*.
- Our method on randomization claims that while approximate retrieval is acceptable only for sake of efficiency, we can further exploit approximate retrieval to provide impressive trade-offs between accuracy and diversity. With this intuition, we design hash functions that are geared to retrieve approximate nearest neighbors in sub-linear time and superior diversity. We formulate the classical multi-label annotation problem (of images and text) in this setting and demonstrate the scalability and diversity in our solution. Our method on learned metrics investigated the problem of image retrieval more closely and shows that appropriate metrics can be learnt to achieve diversity with respect to viewpoint of the camera, time of day, and camera zoom.
- We also define a measure to evaluate diversity in different retrieval tasks of both images and web pages. In particular, we demonstrate effectiveness of our approach in four tasks: Image Category Retrieval, Multi-label Classification, Image Tagging, and Instance based Image Retrieval.

The thesis has been laid out in the following fashion

- In chapter 2, we present motivation, definitions and background information that are essential to understand the rest of the thesis. We formulate the diverse retrieval problem, by highlighting the underlying assumptions as well as inherent differences with the existing formulations. We also describe the use of randomization as a resource and the role of metric design to find appropriate metrics in the instance retrieval.
- In chapter 3, we present our robust algorithm to demonstrate that accuracy, diversity and efficiency can be achieved simultaneously with the help of locality sensitive hash functions. We analysis the algorithmic and statistical aspects of the proposed method with respect all of the above three performance criteria. We show an empirical study in different retrieval scenarios: a) Image Category Retrieval, b) Multi-label Classification, and c) Image Tagging. We also describe basic problems in each diversity task as well as evaluation methodologies and performance measurements.
- In chapter 4, we present a method that is based on learning metrics for the instance based image retrieval. We demonstrate that learning appropriate distance metrics can be effective to retrieve diverse set of images with respect to low-level variations like viewpoint, time of day and camera zoom. Finally, in chapter 5, we draw conclusions from this thesis, and describe some of the future work that follows from this thesis.

Chapter 2

Optimizing Relevance and Diversity

> *The classical Probability Ranking Principle (PRP) forms the theoretical basis for probabilistic Information Retrieval (IR) models, which are dominating IR theory since about 20 years. However, the assumptions underlying the PRP often do not hold, and its view is too narrow. The PRP assumes that documents are relevant independently of one another, so it is not suitable for optimization of novelty or diversity rankings.* - An Analysis of NP-Completeness in Novelty and Diversity Ranking [15, 29]

In this chapter, we define the diverse retrieval problem as an optimization problem. Additionally, we also discuss the inherent difficulty of solving this optimization problem in general. Further, we discuss the key ideas that are used in the design of our alogirithms.

2.1 Motivation

A typical application of multi-label learning is image/video tagging [14, 96, 102], where the goal is to tag a given image with all the relevant concepts/labels. Other examples of multi-label instance classification include bid phrase recommendation [1], categorization of Wikipedia articles etc. In all these applications, the query is typically an instance (*e.g.*, images, text articles) and the goal is to find the most relevant labels (*e.g.,* objects, topics). Moreover, one would like the labels to be diverse. For instance, for a given image of a lab, the appropriate tags might be chair, table, carpet, fan etc. In addition to the above requirement of accurate prediction of the positive labels (tags), we also require the obtained set of positive labels (tags) to be *diverse*. That is, for an image of a lab, we would prefer tags like {table, fan, carpet}, rather than tags like {long table, short table, chair}. However, the given labels are just some names and we typically do not have any features for the labels. Moreover, most of the existing multi-label algorithms run in time linear in the number of labels which renders them infeasible for several real-time tasks [103, 108]; exceptions include random forest based method [1, 73], however, it is not clear how to extend these methods to retrieve diverse set of labels.

Figure 2.1: A single metric is not sufficient when the feature space is heterogeneous i.e., diverse in visual content: a) Visually similar image retrieved using simple BOW based features. b) Retrieving semantically similar images requires computing higher order features such as people, beach etc.

In recent years, content-based image retrieval (CBIR) models [90] have been studied extensively focusing on retrieving images similar to a query or set of queries. These features are automatically extracted from images to compute the similarity between a query and images in database. Existing CBIR applications are heavily reliant on appearance based features like Bag-of-Words BOW [60] indexed by scalable data structures like vocabulary trees, that produce near identical or duplicate results in comparison to the query image. Most of the other CBIR systems use low-level image features, such as color, texture, and shape, to represent the visual content and are heavily specific to the tasks at hand [80]. Therefore, such methods offer limited performance to the capture the perceived image similarity observed by humans.

Figure 2.1 illustrates this case in detail. In the query (a), we notice that the retrieved images can be obtained from simple low-level image features like color, shape or the standard appearance based SIFT [60] features. However, in the query (b), finding semantically similar images is a bit difficult. The low-level features are not suitable for such task and computing the higher-order features like people, beach, water, etc. creates additional overhead in the retrieval process. This may not be appreciated in many application areas, especially, where a diverse set of images are needed. Consequently, there is a need to allow for a generic similarity measures than the pre-defined metrics applied to BOW models or the low-level visual features.

Addressing these shortcomings, we present our investigation on developing efficient algorithms that appreciates better diversity in the retrieval. In particular, we demonstrate that randomization can be a useful resource for efficiently retrieving diverse images, and visual information can be useful to learn metrics that promoted diversity in images with respect to the geometric and illuminance properties. Though the work presented in this thesis is strongly motivated from the image domain, we emphasize that these are equally applicable to other domains like text.

2.2 Problem Formulation

Since, relevance and diversity are relative to users, systems, time, viewpoint, tasks, session state, and other contextual variables, there are several ways to formulate the diverse retrieval problem [21, 33, 52, 76, 109]. In this thesis, we define the diverse retrieval problem in the following manner: Let $\mathcal{X} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a set of data points in $\mathbb{R}^d$ and a query point $q \in \mathbb{R}^d$, the goal is two-fold:

- Retrieve a set of points $\mathcal{R}_q = \{x_{i_1}, \ldots, x_{i_k}\}$ such that a majority of their labels correctly predicts the label of q.
- The set of retrieved points $\mathcal{R}_q$ is "diverse".

Note that, in this work we are only interested in finding k points that are relevant to the query. We formally start with the two definitions that are empirically successful and are widely used measures for similarity and diversity in the context of retrieval:

Definition 2.2.1. *For a given two points, say x and y,* **dis-similarity** *is defined as the distance between the two points, i.e.,*

$$DisSim(x, y) = \|x - y\|_2^2$$

Definition 2.2.2. *For a given set of points,* **diversity** *is defined as the average pairwise distance between the points of the set, i.e.,*

$$Div(\mathcal{R}_q) = \sum_{a,b} \|x_{i_a} - x_{i_b}\|_2^2$$

With the above definitions, our goal is to find a subset of k points which are both relevant to the query and diversified among themselves. Although, it is not quite clear on how relevance and diversity should be combined, we adopt a reminiscent [57] of the general paradigm in machine learning of combining loss functions that measures quality(e.g., training error, prior, or "relevance") and a regularization term that encourages desirable properties (e.g. smoothness, sparsity, or "diversity"). To this end, we define the following optimization problem.

$$\begin{aligned} &\min \ \lambda \Sigma_{i=1}^{n} \alpha_i \|q - x_i\|^2 - (1 - \lambda) \Sigma_{ij} \alpha_i \alpha_j \|x_i - x_j\|^2 \\ &\text{s.t.} \ \ \Sigma_{i=1}^{n} \alpha_i = k; \forall i \in \{1, \ldots n\} \alpha_i \in \{0, 1\} \end{aligned} \tag{2.1}$$

In the equation, $\lambda \in [0, 1]$ is a regularization parameter that defines the trade-off between the two terms, and α_i takes the value 1 if x_i is present in the result and 0 if it is not included in the retrieved result. Intuitively, the first term measures the overall relevance of the retrieved set with respect to the query. The second term measures the similarity among the retrieved points. That is, it penalizes the selection of multiple relevant points that are very similar to each other.

By including this term in the objective function, we seek to find a set of points that are relevant to the query, but also dissimilar to each other. Without loss of generality, we assume that x_i, q are normalized to unit norm, and with some simple substitutions like $\alpha = [\alpha_1, \dots \alpha_n]$, $c = -[q^T x_1, \dots, q^T x_n]$, G be gram matrix with $G_{ij} = x_i^T x_j$, the above objective is equivalent to

$$\begin{aligned} &\min \ \lambda c^T \alpha + \alpha^T G \alpha \\ &\text{s.t.} \ \ \alpha^T 1 = k; \alpha \in \{0,1\}^n \end{aligned} \tag{2.2}$$

From now on, we refer to the diverse retrieval problem in the form of the combinatorial optimization problem in Eq.(2.2). Finding optimal solution for the quadratic integer program in Eq.(2.2) is NP-hard [104]. Therefore, approximation algorithms have been derived by several works. Below, we comprehend the inherent challenges of using the existing solutions which are crucial for the large-scale retrieval applications.

2.3 Key Challenges

There can be several perspectives that one can adopt while trying to solve the optimization problem in Eq.(2.2). The three major perspectives of central importance in this thesis are as follows:

2.3.1 Algorithmic Perspectives

From an algorithmic perspective, the relevant question is: *how long does it take to compute α?* The answer to this question is that, usually QP relaxations [45, 78] (which are often called linear relaxations), where integer constraints are relaxed to interval constraints, are efficiently solvable [49]. With the QP relaxations, we first remove the integrality constraint on the variables i.e., allow variables to take on non-integral values to obtain a quadratic optimization program in Eq.(2.3).

$$\begin{aligned} &\min \ \lambda c^T \alpha + \alpha^T G \alpha \\ &\text{s.t.} \ \ \alpha^T 1 = k; 0 \leq \alpha \leq 1 \end{aligned} \tag{2.3}$$

A variety of methods can be used to find solution to quadratic program in Eq.(2.3). In practice commonly used methods include: interior-point methods [105], active set [66], augmented lagrangian [24], extensions to simplex algorithm [66], etc. Note that the optimal solution to the relaxed problem is not necessarily integral. Therefore, we select the top k values from the fractional solution and report it as the integral feasible solution to Eq.(2.2). Although, all these methods yield a good solution to Eq.(2.2) i.e., obtains accurate and diverse retrieval, for large datasets solving the QP Relaxation is much more time consuming than the existing greedy solutions (*see Table 3.1*).

Thus, a natural question is how to solve this problem either exactly or approximately that suits for all practical purposes? From an algorithmic perspective, one is often interested to show that one can obtain solutions that are approximately as good as the exact solution for the input at hand, in less time than would be required to compute an exact solution for the data at hand.

2.3.2 Statistical Perspectives

From a statistical perspective, the relevant question is: *when is solving this problem right thing to do?* The answer to this question is that, for a class of sub-modular functions which combine two terms like in Eq.(2.2), there exists a simple greedy algorithm [50, 57] who's solution is guaranteed to be almost as good as the optimal solution. For instance, the work in [39] finds a near optimal greedy solution with provable guarantees when the relevance and similarity functions take only non-negative values. That is, their objective function exhibits the diminishing returns property, including sub-modularity, monotonicity, etc. However, our diverse retrieval problem does not make such assumptions.

Thus, a natural question is what to do when the assumptions underlying the problem are not satisfied or are only imperfectly satisfied? From a statistical perspective, one is often more interested in how well a procedure performs relative to the hypothesized model than how well it performs on the particular data set at hand.

To summarize, the existing greedy approaches [13, 26, 39, 110] which have constant factor guarantee of optimality under certain assumptions and the QP relaxation method which are efficiently solvable suffer from three drawbacks: a) Running time of the algorithms is very high as it is required to recover several exact nearest neighbors. b) The obtained points might all be from a very small region of the feature space and hence the diversity of the selected set might not be large. c) Computation of the gram matrix may require an unreasonably large amount of memory overhead for large datasets.

Since, most of the existing methods address only one or two of the above mentioned requirements, it is of greatest interest to look for computationally efficient solutions for the diverse retrieval problem in Eq.(2.2). In this thesis, we develop a simple approach that considers both the algorithmic perspective and statistical perspectives based on randomization and show its effectiveness to meet all of the above requirements simultaneously.

2.3.3 Visual Perspectives

From a visual perspective, the relevant question is: *what is right choice for the similarity measures?* The answer to this question is that, for a class of distance functions, metric learning methods [8, 7, 9, 95] utilize the prior information in the form of pairwise constraints to improve the quality of clusters. In the case of instance based image retrieval, it is highly unlikely that a similarity measure used in Def.(2.2.1) can reflect the true underlying relationships between the images. Figure 2.1 demonstrates that distinction between visual relevance and semantic relevance is difficult when the feature space is heterogeneous.

Thus, a natural question is how to adopt to different aspects of images that are characterized by the physical properties that do not provide any clue on a function that is to be approximated? From a learning perspective, one is often interested to adopt to such properties with respect to structure of the data at hand.

Since, a variety of visual information exists among the images, it is of greatest interest to look for solutions that rely on the prior knowledge of the retrieval task. In this thesis, we develop a simple

approach that learns distance metric and effectively improves diversity in the retrieval with respect to geometric and illuminance qualities of the images.

2.4 Key Ideas

Before describing our approach, we will provide a brief overview of randomized algorithms based on locality sensitive hashing and metric learning algorithms that are related to the main results of this thesis.

2.4.1 Randomization as a resource

Randomization has had a long history in scientific applications [62]. Randomized algorithms [5, 10, 38, 47, 63, 79] for machine learning problems have received a great deal of attention, since several of them can be formulated as discrete optimization problems which are computationally intractable (NP-hard or worse). Apart from the computational hardness of the problem, randomized algorithms have become a very useful in practical tool for two phrases, namely *efficiency* and *proven approximation guarantees*.

Incorporating randomness into machine learning algorithms is not a brand new idea, and they are quite commonly used for training models: a) *Randomization before model induction* methods include Sample randomization for bootstrap sampling [89]; Feature randomization for Randomized Trees and Random Subspace [40]; Data perturbation for Output Smearing and Random Projection [106]; b) *Randomization during model induction* include Partial-random test selection for Tree Randomization and Random Forests [12]; Complete-random test selection for Random Decision Trees [74]. More recently, the need for such methods are strongly motivated by problems in large-scale data analysis applications.

For example, genetics applications [2, 69] consist of large matrices to encode information about the disease genes (i.e., they can be used to perform classification into sick and not sick) as well as population histories (i.e., they can be used to infer properties about population genetics and human evolutionary history). To give a sense of the sizes involved, then it is of size roughly 400 people by 10^6 SNPs, although more recent technological developments [18, 32] have increased the number of SNPs to well into the millions and the number of people to the thousands and tens-of-thousands.

Another recent application demands efficient randomized scheme to perform similarity search on a dataset of >1 billion tweets, and support high-throughput streaming of new data, with hundreds of millions of new tweets per day and achieve query times of 1-2.5 milliseconds. An efficient parallel randomized algorithm pLSH [91] can scale to a large number of nodes to help us analyze very large streaming datasets in real-time. Figure 2.2 shows that approximate algorithms are capable of handling large amounts of streaming data while delivering very high query performance. For applications that require real-time responses, trading off a small hit in accuracy for faster speed of processing is reasonable and acceptable may even be inevitable.

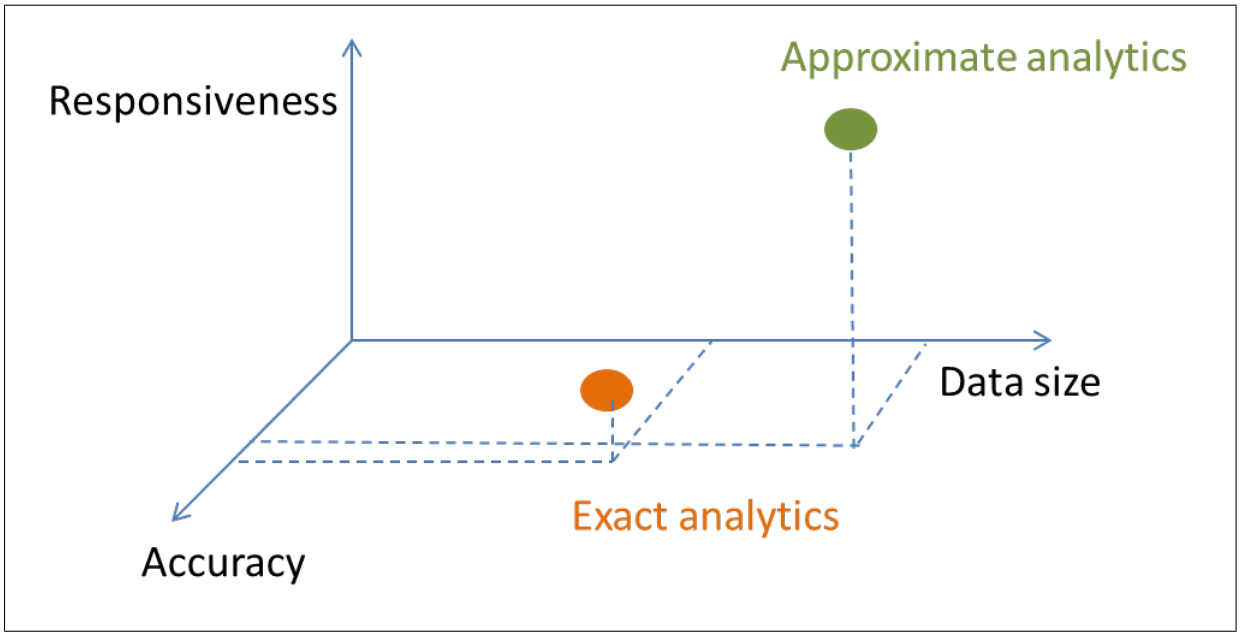


Figure 2.2: Trading off a small hit in accuracy for faster speed of processing is reasonable and acceptable. For applications that require real-time responses, this trade-off may even be inevitable [91].

Depending on the size of the data and the problem under consideration, randomized algorithms can be useful in one or more of several ways. In this work, we argue that *randomization after model induction* can be a useful resource to improve diversity in the retrieval. We use a simple algorithm that enables fast retrieval by using locality sensitive hash functions and show its subsequent relevance to improve diversity in retrieval with very large databases.

2.4.2 Locality Sensitive Hashing

Motivated by the dominance based redundancy elimination property of skylines, an approximate k-NN based skyline algorithm is proposed in [92]. They adapt a multidimensional indexing scheme that uses a B+ tree [44, 93] to index each individual attribute or dimension. The most similar samples suggested by each attribute then form the unique candidate list of skyline points. These candidate similar objects are then processed to eliminate conceptually redundant samples. Together, the result is an efficiently computed skyline of the database with respect to the query, resulting in a diverse set of similar results.

However, it is also observed in some scenarios especially when there are large number of attributes it becomes computationally expensive, and skylines may not be useful and advantageous. This distinction between non-redundancy and diversity manifests itself more in dense neighborhoods especially where the attributes belong to a high precision continuous space, like real numbered visual features.

In our work, we use the Locality-sensitive Hashing (LSH) [17, 34, 41], a popular approach for similarity search on high-dimensional data. High level idea behind LSH based methods is to use hash

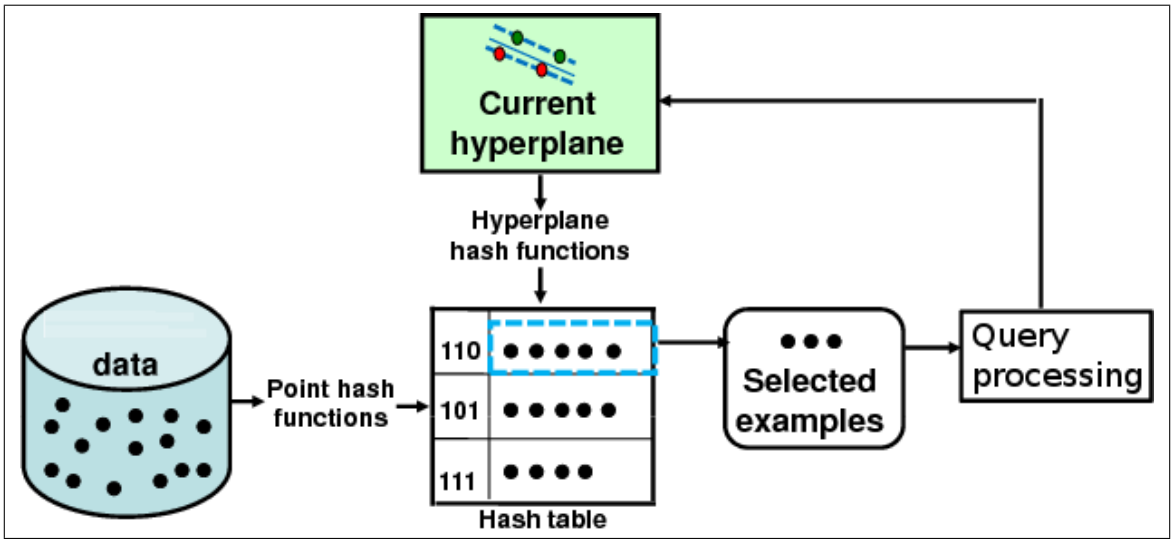


Figure 2.3: Off-line Step: Hash unlabelled data into table. On-line Step: Hash current classifier as "query" to directly report near by points. For more details refer to [43]

functions to map similar points to the same hash buckets, so that only a subset of the database must be searched. That is, hash the points using several hash functions so as to ensure that, for each function, the probability of collision is much higher for objects which are close to each other than for those which are far apart. Then, one can retrieve (approximate) neighbors by hashing the query point and reporting all elements stored in buckets containing that point. Figure 2.3 demonstrates the idea of hashing queries to retrieve similar points.

The LSH algorithm has several variants, depending on the underlying distance functions. Existing LSH functions can accommodate the Hamming distance [41], L_p norms [22], and inner products [17], and such functions have been explored previously in the vision community [37, 65, 84]. Data-dependent variants of LSH have been proposed: the authors of [31] select partitions based on where data points are concentrated, while in [84] boosting is used to select feature dimensions that are most indicative of similarity in the parameter space. The authors of [42] proposed a method for fast approximate similarity search with learned Mahalanobis metrics.

While earlier approaches considered locality sensitive hashing to be acceptable only for the sake of efficiency, we argue that one can further exploit the randomness to provide impressive trade-offs between accuracy and diversity in the retrieval. To this end, we propose a locality sensitive hashing based algorithm that guarantees approximate nearest neighbor (ANN) retrieval in sub-linear time and with superior diversity. Our method retrieves points that are sampled uniformly at random to ensure diversity in the retrieval while maintaining reasonable number of relevant ones.

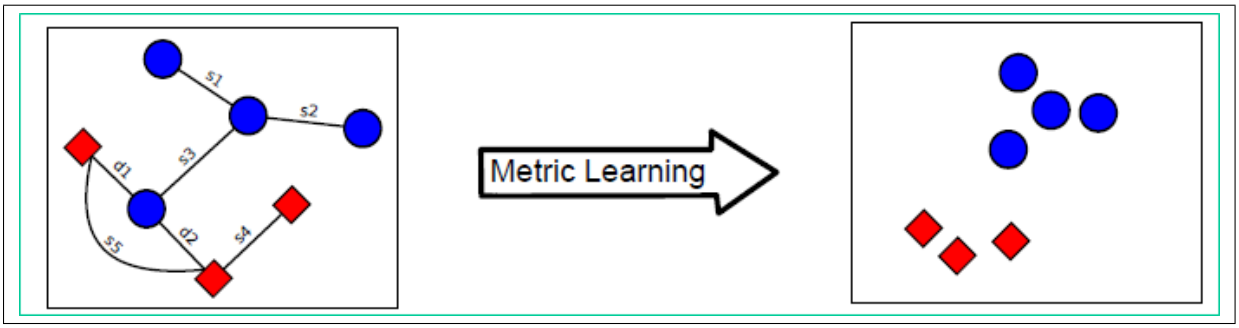


Figure 2.4: Metric Learning maps similar samples close together and dissimilar samples far apart as measured by the learned Mahalanobis metric.

2.4.3 Metric Design by Learning

The key to measure accurate visual similarity between images is to find appropriate distance metric for the given task at hand. In the definition 2.2.1, we use a pre-defined distance metric for image similarity measurement. A more general form is given by Mahalanobis distance,

$$d_A(x, y) = (x - y)^T A(x - y) \tag{2.4}$$

where, A is symmetric, positive semi-definite matrix. Note that if $A = I$, the above equation is equivalent to the def.(2.2.1) i.e., Euclidean distance metric. As discussed in the section 2.3.3, our goal is to learn distance metrics A from the image specific information to encourage diversity.

Metric learning [100] is the process of adapting a metric according to side-information about the similarity or dissimilarity of some known data points. In this thesis, we use a popular metric learning, Information theoretic metric learning (ITML) [23]. ITML algorithm uses an information-theoretic cost model which iteratively enforces pairwise similar/dissimilarity constraints, yielding a learned Mahalanobis distance metric, A. ITML utilizes the prior knowledge about the inter-point distances under simple similar and dissimilar constraints in the initial feature space.

Note that a pairwise distance computation by Eq.(2.4) can also be realized by first performing a linear transformation $\mathcal{X} \rightarrow \mathcal{T} = A^{\frac{1}{2}}\mathcal{X}$ and by computing the L^2 distance for the pair in $\mathcal{T}$. This linear transformation makes similar data points in $\mathcal{X}$ closer together and dissimilar data points farther apart in $\mathcal{T}$, and yields more computationally efficient pairwise computation.

Adopting this property, we treat the ITML's result A as a post feature transformation and show its effectiveness for improving the performance in the image retrieval. In particular, in section 4.2, we show that such feature transformation can result in improvements in diversity with respect to the low-level variations in images like geometric and illuminance qualities.

Chapter 3

Diversity using Randomization

> *In the same way as space and time are valuable resources available to be used judicuosly by algorithms, it has been discovered that exploiting* ***randomness as an algorithmic resource*** *inside the algorithm can lead to better algorithms.* - Foundations and Trends in Machine Learning Series [62]

In this chapter, we present our LSH based approach and derive a theoretically sound result to support the intuitiveness and reliability of using LSH functions to improve diversity in retrieval. We modify the standard hash functions to present a fast and efficient algorithm for diverse retrieval. We also extend our method to the problem of multi-label classification, where the goal is to output a diverse and accurate set of labels for documents in real-time. We analyse the proposed algorithm in detail and illustrate its effectiveness over existing approaches. Further, we apply our approach in different retrieval settings and demonstrates its efficacy.

3.1 Diversity with Hash Functions

To find nearest neighbors, the basic LSH algorithm concatenates a number of functions $h \in \mathcal{H}$ into one hash function $g \in \mathcal{G}$. Informally, we say that $\mathcal{H}$ is *locality-sensitive* if for any two points a and b, the probability of a and b collide under a random choice of hash function depends only on the distance between a and b. Several such families are known in the literature, see [3] for an overview.

Definition 3.1.1. *(Locality-sensitive hashing): A family of hash functions $\mathcal{H} : R^d \rightarrow \{0, 1\}$ is called (r, ϵ, p, q)-sensitive if for any $a, b \in R^d$*

$$\begin{cases} Pr_{h\in\mathcal{H}}[h(a) = h(b)] \geq p, & \text{if } d(a, b) \leq r \\ Pr_{h\in\mathcal{H}}[h(a) = h(b)] \leq q, & \text{if } d(a, b) \geq (1 + \epsilon)r \end{cases}$$

Here, $\epsilon > 0$ is an arbitrary constant, $p > q$ and $d(., .)$ is some distance function.

In this thesis, we use ℓ_2 norm as the distance function and adopt the following hash function,

$$h(a) = sign(r \cdot a) \tag{3.1}$$

where $r \sim \mathcal{N}(0, I)$. It is well known that $h(a)$ is a LSH function w.r.t ℓ_2 norm and it is shown to satisfy the following:

$$Pr(h(a) \neq h(b)) = \frac{1}{\pi} \cos^{-1}\left(\frac{a \cdot b}{\|a\|_2 \|b\|_2}\right). \tag{3.2}$$

Our approach is based on the following high-level idea: perform *randomized approximate* nearest neighbor search for q which selects points randomly from a small disk around q. As we show later, locality sensitive hashing with standard hash functions actually possess such a quality. Hence, the retrieved set would not only be accurate (i.e. has small distance to q) but also diverse as the points are selected randomly from the neighborhood of q.

Algorithm 1: LSH with random vectors for Diversity (LSH-Div)

Input: $\mathcal{X} = \{x_1 \ldots, x_n\}$, where $x_i \in R^d$, a query $q \in \mathbb{R}^d$ and k an integer.

1 **Preprocessing**: For each $i \in [1 \ldots L]$, construct a hash function, $g_i = [h_{1,i}, \ldots, h_{l,i}]$, where $h_{1,i}, \ldots, h_{l,i}$ are chosen at random from $\mathcal{H}$. Hash all points in $\mathcal{X}$ to the i^{th} hash table using the function g_i

2 $R \leftarrow \phi$

3 **for** $i \leftarrow 1$ ***to*** L **do**

4 Perform a hash of the query $g_i(q)$

5 Retrieve points from i^{th} hash table & append to $\mathcal{R}_q$

6 $\mathcal{S}_q \leftarrow \phi$

7 **for** $i \leftarrow 1$ ***to*** k **do**

8 $r^* \leftarrow \text{argmin}_{(r \in \mathcal{R}_q)}(\lambda \|q - r\|^2 + \frac{1}{i} \Sigma_{s \in \mathcal{S}_q} \|r - s\|^2)$

9 $\mathcal{R}_q \leftarrow \mathcal{R}_q \setminus r^*$

10 $\mathcal{S}_q \leftarrow \mathcal{S}_q \cup r^*$

Output: $\mathcal{S}_q$, k diverse set of points

The algorithm executes in two phases: i) perform search through the hash tables, line(2-4), to report the approximate nearest neighbors, $R_q \subset \mathcal{X}$ and ii) perform k iterations, line(6-9), to report a diverse set of points, $S_q \subset R_q$. Throughout the algorithm, several variables are used to maintain the trade-off between the accuracy and diversity of the retrieved points. The essential control variables that direct the behaviour of the algorithm are: i) the number of points retrieved from hashing, $|R_q|$ and ii) the number of diverse set of points to be reported, k. Here, R_q can be controlled at the design of hash function, i.e., the number of matches to the query is proportional to $n^{\frac{1}{1+\epsilon}}$. Therefore, line 7 is critical for the efficiency of the algorithm, since it is an expensive computation, especially when $|R_q|$ is very big, or k is large. More details of our approach are described in section 3.3.

3.1.1 Diversity in Randomized Hashing

The above mentioned LSH function is unbiased towards any particular direction, i.e., $Pr(h(q) \neq h(a))$ is dependent only on $\|q - a\|_2$ (assuming q, a are both normalized to unit norm vectors). But, depending on a sample hyper-plane $r \in \mathbb{R}^d$, a hash function can be biased towards one or the other

direction, hence preferring points from a particular region. Interestingly, we show that if the number of hash bits is large, then all the directions are sampled uniformly and hence the retrieved points are sampled uniformly from all the directions. *That is, the retrieval is not biased towards any particular region of the space.* We formalize the above observation in the following lemma.

Definition 3.1.2. *(***Hoeffding inequality** *[11, 61]): Let $Z_1, \ldots, Z_n$ be n i.i.d. random variables with $f(Z) \in [a, b]$. Then with probability at least $1 - \delta$ we have*

$$P[\|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E(f(Z))\|] \leq (b-a)\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}$$

Lemma 3.1.1. *Let $q \in \mathbb{R}^d$ and let $\mathcal{R}_q = \{x_1, \ldots, x_m\}$ be unit vectors such that $\|q - x_i\|_2 = \|q - x_j\|_2 = r$, $\forall i, j$. Let $p = \frac{1}{\pi}\cos^{-1}(1 - r^2/2)$. Also, let $r_1, \ldots, r_\ell \sim \mathcal{N}(0, I)$ be ℓ random vectors. Define hash bits $g(x) = [h_1(x) \ldots h_\ell(x)] \in \{0, 1\}^{1\times\ell}$, where hash functions $h_b(x) = sign(r_b \cdot x), 1 \leq b \leq \ell$. Then, the following holds $\forall i$:*

$$p - \sqrt{\frac{\log(\frac{2}{\delta})}{2l}} \leq \frac{1}{l}||g(q) - g(x_i)||_1 \leq p + \sqrt{\frac{\log(\frac{2}{\delta})}{2l}}$$

That is, if $\sqrt{l} \gg 1/p$, then hash-bits of the query q are almost equi-distant to the hash-bits of each x_i.

Proof. Consider random variable $Z_{ib},\ 1 \leq i \leq m,\ 1 \leq b \leq \ell$ where $Z_{ib} = 1$ if $h_b(q) \neq h_b(x_i)$ and 0 otherwise. Note that Z_{ib} is a Bernoulli random variable with probability p. Also, $Z_{ib},\ \forall 1 \leq b \leq \ell$ are all independent for a fixed i. Hence, applying Hoeffding's inequality, we obtain the required result. □

Note that the above lemma shows that if $x_1, \ldots, x_m$ are all at distance r from a given query q then their respective hash bits are also at a similar distance to the hash bits of q. *That is, assuming randomization selection of the candidates from a hash bucket, probability of selecting any x_i is almost the same. That is, the points selected are nearly uniform at random and are diverse.*

3.1.2 Compact Randomized Hashing

In the previous section, we obtained hash functions by selecting hyper-planes r from a normal distribution. The conventional LSH approach considers only random projections. Naturally, by doing random projection, we will lose some accuracy. But we can easily fix this problem by doing multiple rounds of random projections. However, we need to perform a large number of projections (i.e. hash functions in the LSH setting) to increase the probability that similar points are mapped to similar hash codes. A fundamental result of Johnson and Lindenstrauss Theorem [46] says that $O(\frac{\ln N}{\epsilon^2})$[1] random projections are needed to preserve the distance between any two pair of points, where ϵ is the relative error.

[1]Note that in LSH schemes, number of matches to the query is proportional to $N^{\frac{1}{1+\epsilon}}$, where N is the total number of points in the database.

Therefore, using many random vectors to generate the hash tables (a long codeword), leads to a large storage space and a high computational cost, which would slow down the retrieval procedure. In practice, however, the data lies in a very small dimensional subspace of the ambient dimension and hence a random hyper-plane may not be very informative. Instead, we wish to use more data driven hyper-planes that are more discriminative and separate out neighbors from far-away points. To this end, we obtain the hyper-planes r using principal components of the given data matrix. Principal components are the directions of highest variance of the data and captures the geometry of the dataset accurately. Hence, by using principal components, we hope to reduce the required number of hash bits and hash tables required to obtain the same accuracy in retrieval.

That is, given a data matrix $X \in \mathbb{R}^{d \times n}$ where i-th column of X is given by x_i, we obtain top-α principal components of X using SVD [54]. That is, let $U \in \mathbb{R}^{d \times \alpha}$ be the singular vectors corresponding to the top-α singular values of X. Then, a hash function is given by:

$$h(x) = sign(r^T U^T x) \tag{3.3}$$

where $r \sim \mathcal{N}(0, I)$ is a random α-dimensional hyper-plane.

Algorithm 2: LSH with singular vectors for Diversity (LSH-SDiv)

Input: $\mathcal{X} = \{x_1 \ldots, x_n\}$, where $x_i \in R^d$, a query $q \in \mathbb{R}^d$, k an integer, α number of singular vectors.

1 $[\Lambda; U] = SVD(\mathcal{X}; \alpha)$

2 Construct the hash function $g_i = [h_{1,i}, \ldots, h_{l,i}]$, where $h_{1,i}, \ldots, h_{l,i}$ are randomly chosen α - dimensional hyperplanes according to Eq.(3.3).

3 **Execute lines 1-9 from LSH-Div.**

Output: $\mathcal{S}_q$, k diverse set of points

Many learning based hashing methods [53, 97, 101] are proposed in literature. The simplest of all such approaches is PCA Hashing [99] which chooses the random projections to be the principal directions of the data directly. Our algorithm LSH-SDiv method is different from PCA Hashing in the sense that we still select random directions in the top components. Note that the above hash function has reduced randomness but still preserves the discriminative power by projecting the randomness onto top principal components of X. As shown in Section 3.5, the above hash function provides better nearest neighbor retrieval while recovering more diverse set of neighbors.

3.2 Diverse Multi-label Prediction

Let $\mathcal{X} = \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^d$ and $\mathcal{Y} = \{y_1, \ldots, y_n\}$, where $y_i \in \{-1, 1\}^L$ be L labels associated with the i-th data point. Then, the goal in the standard multi-label learning problem is to predict the label vector y_q accurately for a given query point q. Moreover, in practice, the number of labels L is very large, so we require our prediction time to scale sub-linearly with L. We propose a method that guarantees diverse *and* sub-linear (in the number of labels) time multi-label prediction.

Algorithm 3: LSH based Multi-label Prediction

Input: Train data: $\mathcal{X} = \{x_1, \ldots, x_n\}$, $\mathcal{Y} = \{y_1, \ldots, y_n\}$. Test data: $\mathcal{Q} = \{q_1, \ldots, q_m\}$. Parameters: α, k.

[W, H]=LEML($\mathcal{X}$, $\mathcal{Y}$, k);

$\mathcal{S}_q =$ LSH-SDiv$(W, H^T q, \alpha)$, $\forall q \in \mathcal{Q}$;

$\widehat{y}_q =$ Majority$(\{y_i \text{ s.t. } x_i \in \mathcal{S}_q\})$, $\forall q \in \mathcal{Q}$;

Output: $\widehat{\mathcal{Y}}_Q = \{\widehat{y}_{q_1}, \ldots, \widehat{y}_{q_m}\}$

Our method is based on the LEML method [108] which is an embedding based method. The key idea behind embedding based methods for multi-label learning is to embed *both* the given set of labels as well as the data points into a *common* low-dimensional space. The relevant labels are then recovered by NN retrieval for the given query point (in the embedded space). That is, we embed each label i into a k-dimensional space (say $y_i \in \mathbb{R}^k$) and the given test point is also embedded in the same space (say $x_q \in \mathbb{R}^k$). The relevant labels are obtained by finding y_i's that closest to x_q. Note that as the final prediction reduces to just NN retrieval, we can apply our method to obtain diverse set of labels in sub-linear time.

In particular, LEML learns matrices W, H s.t. given a point q, its predicted labels is given by $y_q = sign(WH^T x)$ where $W \in \mathbb{R}^{L \times k}$ and $H \in R^{d \times k}$ and k is the rank of the parameter matrix WH^T. Typically, $k \ll \min(d, L)$ and hence the method scales linearly in both d and L. For instance, its prediction time is given by $O((d + L) \cdot k)$. However, for several widespread problems, the $O(L)$ prediction time is quite large and makes the method infeasible in practice. Moreover, the obtained labels from this algorithm can all be very highly correlated and might not provide a diverse set of labels which we desire.

We overcome both of the above limitations of the algorithm using the LSH based algorithm introduced in the previous section. We now describe our method in detail. Let $W_1, W_2, \ldots, W_L$ be L data points where $W_i \in \mathbb{R}^{1 \times k}$ is the i-th row of W. Also, let $H^T x$ be a query point for a given x. Note that the task of obtaining α positive labels for given x is equivalent to finding α largest $W_i \cdot (H^T x)$. Hence, the problem is the same as nearest neighbor search with diversity where the data points are given by $\mathcal{W} = \{W_1, W_2, \ldots, W_L\}$ and the query point is given by $q = H^T x$.

We now apply our LSH method (Algorithm 1 and 2) to the above setting to obtain a "diverse" set of labels for the given data point x. Moreover, the LSH Theorem by [34] shows that the time of retrieval is sub-linear in L which is necessary for the approach to scale to a large number of examples. See Algorithm 3 for the pseudo-code of our approach.

3.3 Algorithmic and Statistical Analysis

As discussed above, locality sensitive hashing is a sub-linear time algorithm for near(est) neighbor search that works by using a carefully selected hash function that causes objects or documents that are similar to have a high probability of colliding in a hash bucket. Like most indexing strategies, LSH

consists of two phases: *hash generation*, where the hash tables are constructed and *querying*, where the hash tables are used to look up for points similar to the query. Here, we briefly comment on the algorithmic and statistical aspects which are important for the suggested algorithms in the previous sections.

Hash Generation: In our algorithm, for l specified later, we use a family $\mathcal{G}$ of hash functions $g(x) = (h_1(x), \ldots, h_l(x))$, where $h_i \in H$. For an integer L, the algorithm chooses L functions $g_1, \ldots, g_L$ from $\mathcal{G}$, independently and uniformly at random. The algorithm then creates L hash arrays, one for each function g_j. During preprocessing, the algorithm stores each data point $x \in \mathcal{X}$ into bucket $g_j(x)$ for all $j = 1, \ldots, L$. Since the total number of buckets may be large, the algorithm retains only the non-empty buckets by resorting to standard hashing.

Querying: To answer a query q, the algorithm evaluates $g_1(q), \ldots, g_L(q)$, and looks up the points stored in those buckets in the respective hash arrays. For each point p found in any of the buckets, the algorithm computes the distance from q to p, and reports the point p if the distance is at most r. Different strategies can be adopted to limit the number of points reported to the query q, see [3] for an overview.

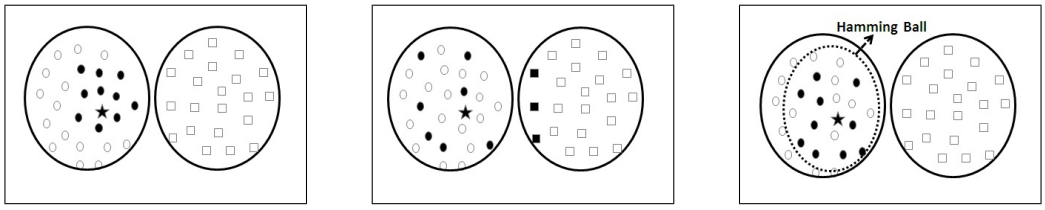
Accuracy: Since, the data structure used by LSH scheme is randomized: the algorithm must output all points within the distance r from q, and can also output some points within the distance $(1 + \epsilon)r$ from q. The algorithm guarantees that each point within the distance r from q is reported with a constant (tunable) probability. The parameters l and L are chosen [41] to satisfy the requirement that a near neighbors are reported with a probability at least $(1 - \delta)$. Note that the correctness probability is defined over the random bits selected by the algorithm, and we do not make any probabilistic assumptions about the dataset.

Diversity: In lemma 3.1.1, if the number of hash bits is large i.e, if $\sqrt{l} \gg 1/p$, then hash-bits of the query q are almost equi-distant to the hash-bits of each point in x_i. Then all the directions are sampled uniformly and hence the retrieved points are uniformly spread in all the directions. Therefore, for reasonable choice of the parameter l, the proposed algorithm obtains diverse set of points, $\mathcal{S}_q$ and has strong probabilistic guarantees for large databases of arbitrary dimensions.

Scalability: The time for evaluating the g_i functions for a query point q is $O(dlL)$ in general. For the angular hash functions chosen in our algorithm, each of the l bits output by a hash function g_i involves computing a dot product of the input vector with a random vector defining a hyperplane. Each dot product can be computed in time proportional to the number of non-zeros ζ rather than d. Thus, the total time is $O(\zeta lL)$. For an interested reader, see that the Theorem 2 of [17] guarantees that L is at most $O(N^{\frac{1}{(1+\epsilon)}})$, where N denotes the total number of points in the database.

3.4 A retrospective and a prospective

Many of the diversification approaches are centered around an optimization problem that is derived from both relevance and diversity criteria (Eq.(2.1)). Most popular among these methods is MMR optimization [13] which recursively builds the result set by choosing the next optimal selection given



(a) k-NN: Accurate, Not diverse (b) Greedy: Not Accurate, diverse (c) ANN: Accurate, diverse

Figure 3.1: Consider a toy dataset with two classes: class A ($\circ$) and class B ($\square$). We show the query point ($\star$) along with ten points ($\bullet, \blacksquare$) retrieved by various methods. In this case, we consider diversity to be the average pairwise distance between the points. a) A conventional similarity search method (e.g: k-NN) chooses points very close to the query and therefore, shows poor in diversity. b) Existing greedy methods offer diversity but might make poor choices by retrieving points from the class B. c) Our method first finds a large set of nearest neighbors within a hamming ball of a certain radii around the query point and then greedily selects points to further improve the diversity.

the previous optimal selections. This greedy strategy intends to retrieve the k diverse points to a query in two steps: First pick a point that is most similar to the query. In the next sequence of steps, the algorithm iteratively selects a point (different from already selected points) that is optimum according to some ad hoc (chosen a-prior) criterion. Essentially, the algorithm enforces relevance scores of the next added point is "near" to the query, and the diversity score to make it be "far" from the current solution. After k iterations, the algorithm retrieves k diverse points. Thus runtime for such greedy search methods is linear in the number of data points, i.e., $O(kdN)$. This is tolerable for small data sets, but it is too inefficient for large ones. The "sole objective" of our research is to design an algorithm for this problem that achieves sub-linear query time. It is clear from previous discussions that our algorithm can do well in selecting accurate and diverse points. Figure 3.1 contrasts our approach with two existing approaches on a toy dataset.

Randomize don't Optimize: Recent works [82, 98] have shown that natural forms of diversification arise via optimization of rank-based relevance criteria such as average precision and reciprocal rank. It is conjectured that optimizing n-$call@k$ metric correlates more strongly with diverse retrieval. More specifically, it is theoretically shown [82] that greedily optimizing expected 1-$call@k$ w.r.t a latent subtopic model of binary relevance leads to a diverse retrieval algorithm that shares many features to the MMR optimization. However, the existing greedy approaches that try to solve the related optimization problem are computationally more expensive than the simple NN, rendering them infeasible for large scale retrieval applications.

Complementary to these methods, our work recommends diversity in retrieval using randomization and not optimization. In our work, instead of finding exact nearest neighbors to a query, we retrieve approximate nearest neighbors that are diverse. Intuitively, our work parallels with these works [82, 98], and generalizes to arbitrary relevance/similarity function. In our findings, we theoretically show that approximate NN retrieval via locality sensitive hashing naturally retrieve points which are diverse.

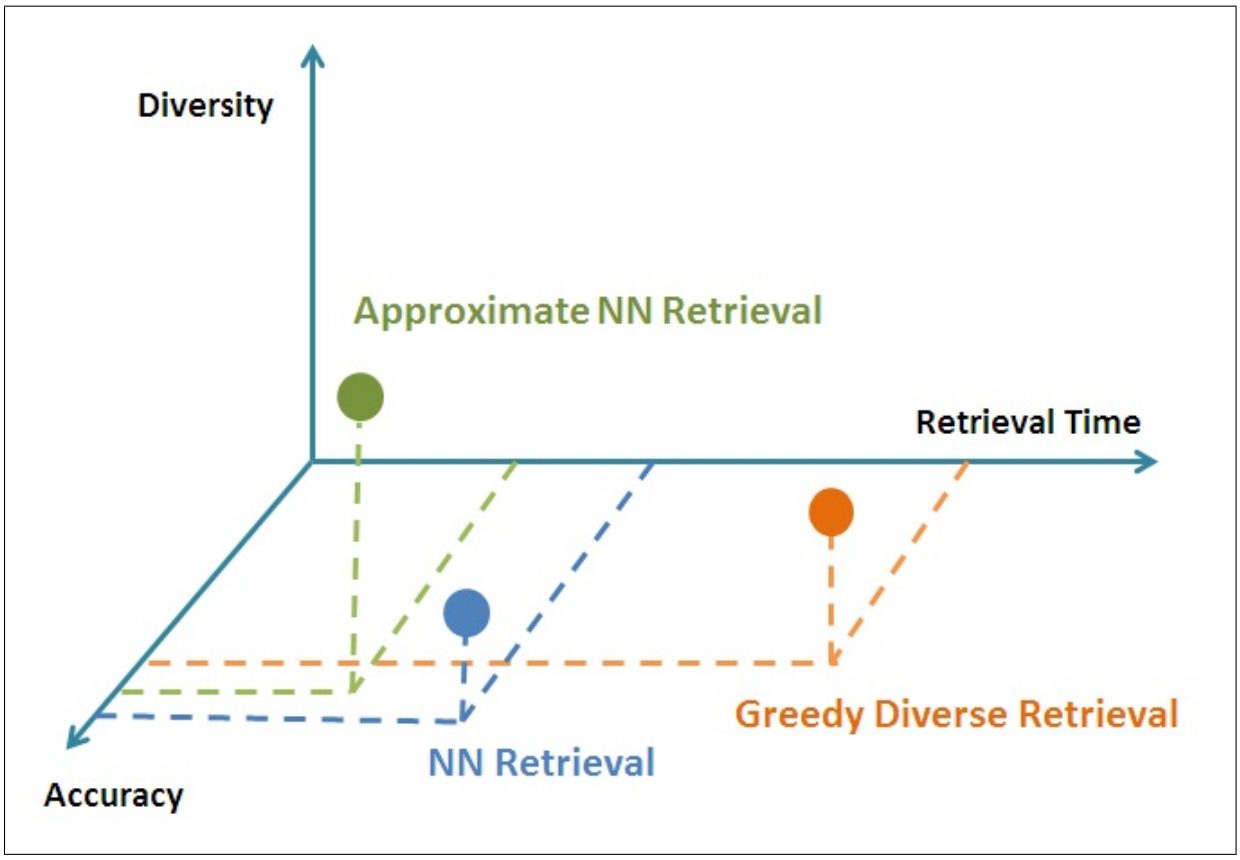


Figure 3.2: Conventional NN method retrieves most accurate results but are very poor at diversity. Existing greedy diverse retrieval methods obtain diverse points with loss in accuracy but are computationally expensive. In our work, we show that approximate algorithms via randomization promise sub-linear retrieval time while trading off a small hit in accuracy and simultaneously improving diversity in the retrieval.

Although the need for an evaluation capability is universally conceded, some commentators have come to feel that the search for a "single" measure of effectiveness is misguided - that there is not now and never can be any one "correct" way of measuring retrieval success. Others regard this attitude as a mere counsel of despair, and continue to look for the "right" way to evaluate system output, in its most essential features at least, and relative to whatever the motive of the system evaluation happens to be. - On Selecting a Measure of Retrieval Effectiveness by Cooper [19, 20]

3.5 Experiments

We demonstrate the effectiveness of our approach in the following three tasks: (a) Image Retrieval (b) Multi-label Classification, and (c) Image Tagging. Our common goal is to demonstrate that our methods can effectively: 1) retrieve accurate results and 2) show high diversity among the retrieved results. Additionally, for very large databases, we also show the efficiency of our approach when compared to the existing diverse retrieval methods.

In the case of image retrieval task, we are interested in retrieving diverse images of a specific category. In our case, each of the image categories have associated subcategories (e.g., *flower* is a category and *lilly* is a subcategory) and we would like to retrieve the relevant (to the category) but diverse images that belong to different sub-categories. The query is represented as a hyperplane that is trained (SVM [85]) off-line to discriminate between positive and negative classes.

Next, we apply our diverse retrieval method to the multi label classification problem; see previous section for more details. Our approach is evaluated on LSHTC[2] dataset containing Wikipedia text documents. Each document is represented with the help of a set of categories or class labels. A document can have multiple labels. Given a test document, we are interested in assigning a set of categories to the document. We model this problem as retrieving a relevant set of labels from a large pool of labels. In this case, we are interested in retrieving labels that match the semantics of the document and also have enough diversity among the labels.

Finally, we consider the image tagging problem which is also a multi label classification problem. In the case of image tagging, we are interested in efficiently predicting a set of relevant tags for a given image, however with diverse semantics to each other.

Evaluation Measures: We now present formal metrics to measure performance of our method on three key aspects of NN retrieval: (i) accuracy (ii) diversity and (iii) efficiency. We characterize the performance in terms of the following measures:

- **Accuracy:** We denote precision at k ($P@k$) as the measure of accuracy of the retrieval. This is the proportion of the relevant instances in the top k retrieved results. In our results, we also

[2] http://lshtc.iit.demokritos.gr/LSHTC3_CALL

report the recall and f-score results when applicable, to compare the methods in terms of multiple measures.

- **Diversity:** For image retrieval, the diversity in the retrieved images is measured using entropy as $D = \frac{\Sigma_{i=1}^{m} s_i \log s_i}{\log m}$, where s_i is the fraction of images of i^{th} subcategory, and m is the number of subcategories for the category of interest. For multi label classification, the relationships between the labels is not a simple tree. It is better captured using a graph and the diversity is then computed using drank [64]. Drank captures the extent to which the labels of the documents belong to multiple categories. For the image tags, diversity is computed with the help of word nets [70]. We use the path similarity which computes the shortest path that connects the senses in the is-a (hypernym/hypnoym) taxonomy.

- **Efficiency:** Given a query, we consider retrieval time to be the time between posing a query and retrieving images/labels from the database. For LSH based methods, we first load all the LSH hash tables of the database into the main memory and then retrieve images/labels from the database. Since, the hash tables are processed off-line, we do not consider the time spent to load the hash tables into the retrieval time. All the retrieval times are based on a Linux machine with Intel E5-2640 processor(s) with 96GB RAM.

Combining Accuracy and Diversity: A ranked list built from a collection that does not cover multiple subtopics cannot be diversified; neither can a ranked list that contains no relevant documents. To ensure that we are assessing systems fairly, the evaluation measure should take into account both relevance and diversity. Trade-offs between accuracy and efficiency in NN retrieval have been studied well in the past [6, 43, 77, 107]. Many methods compromise on the accuracy for better efficiency. Similarly, emphasizing higher diversity may also lead to poor accuracy and hence, we want to formalize a metric that captures the trade-off between diversity and accuracy. To this end, we use (per data point) harmonic mean of accuracy and diversity as overall score for a given method (similar to f-score providing a trade off between precision and recall). That is, $h\text{-}score(\mathcal{A}) = \sum_i \frac{2 \cdot Acc(x_i) \cdot Diversity(x_i)}{Acc(x_i) + Diversity(x_i)}$, where $\mathcal{A}$ is a given algorithm and x_i's are given test points. In all of our experiments, parameters are chosen by cross validation such that the overall h-score is maximized.

3.5.1 Image Category Retrieval

For the image category retrieval, we consider a set of $42K$ images from ImageNet database [25] with 7 synsets (categories) (namely *animal, bottle, flower, furniture, geography, music, vehicle*) with five subtopics for each. Images are represented as a bag of visual words histogram with a vocabulary size of $48K$ over the densely extracted SIFT vectors. For each categorical query, we train an SVM hyperplane using LIBLINEAR [27]. Since, there are only seven categories in our dataset, for each category we created 50 queries by randomly sampling 10% of the images. After creating the queries, we are left with $35K$ images which we use for the retrieval task. We report the quantitative results in

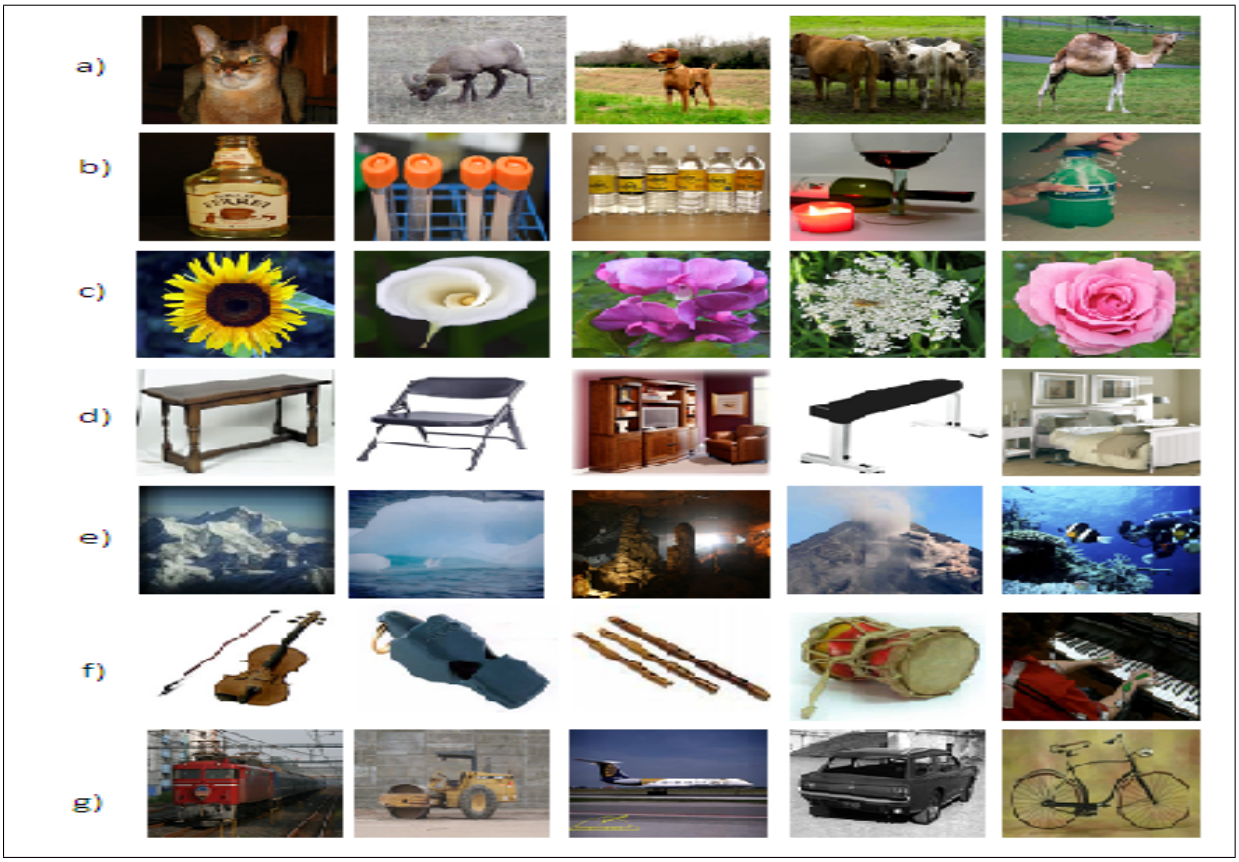


Figure 3.3: Seven concepts from ImageNet database: *a) animal b) bottle c) flower d) furniture e) geography f) music and g) vehicle*. Each row shows example images of a category with five different sub-topic images.

Table 3.1 by the mean performance of all 350 queries. A few qualitative results on this dataset are shown in Figure 3.4.

We conducted two sets of experiments, 1) Retrieval without using hash functions and 2) Retrieval using hash functions, to evaluate the effectiveness of our proposed method. In the first set of experiments, we directly apply the existing diverse retrieval methods on the complete dataset. In the second set of experiments, we first select a candidate set of points by using the hash functions and then apply one of these methods to retrieve the images.

We hypothesize that using hash functions in combination with any of the diverse retrieval methods will improve the diversity and the overall performance (h-score) with significant speed-ups. To validate our hypothesis, we evaluate various diverse retrieval methods in combination with our hash functions as described in Algorithm 1 and Algorithm 2. It can be noted that lines 6-10 in Algorithm 1 can be replaced with various retrieval methods and can be compared against the methods without hash functions. In particular, we show the comparison with the following retrieval methods: the k-nearest neighbor (NN), the QP-Rel method and the diverse retrieval methods like Backward selection (Re-rank), Greedy [39],

Table 3.1: We show the performance of various diverse retrieval methods on the ImageNet dataset. We evaluate the performance in terms of precision(P), sub-topic recall(SR) and Diversity(D) measures at top-10, top-20 and top-30 retrieved images. Numbers in **bold** indicate the top performers. **NH** *corresponds to the method without using any hash function.* Notice that for all methods, *except Greedy*, LSH-Div and LSH-SDiv hash functions consistently show better performance in terms of h-score than the method with NH. Interestingly, we also have the top performers best in terms of retrieval time.

		precision at 10					**precision at 20**					**precision at 30**				
Method	**Hash Function**	P	SR	D	h	time (sec)	P	SR	D	h	time (sec)	P	SR	D	h	time (sec)
NN	NH	1.00	0.60	0.53	0.66	0.621	0.99	0.72	0.60	0.73	0.721	0.99	0.79	0.65	0.77	0.845
	LSH-Div	0.97	0.79	0.76	**0.84**	**0.112**	0.93	0.93	0.86	**0.89**	**0.137**	0.89	0.98	0.91	0.90	0.179
	LSH-SDiv	0.98	0.76	0.73	0.83	0.181	0.95	0.89	0.85	0.89	0.183	0.92	0.95	0.89	**0.90**	**0.106**
Rerank	NH	1.00	0.73	0.69	0.81	0.804	0.99	0.79	0.70	0.81	0.793	0.99	0.88	0.77	0.86	0.901
	LSH-Div	0.93	0.80	0.76	0.83	0.142	0.92	0.93	0.86	0.88	0.146	0.87	0.98	0.90	0.88	0.214
	LSH-SDiv	0.95	0.79	0.76	**0.84**	**0.154**	0.94	0.91	0.85	**0.89**	**0.179**	0.90	0.95	0.88	**0.89**	**0.203**
Greedy [26]	NH	0.95	0.75	0.71	0.80	5.686	0.98	0.86	0.77	**0.85**	11.193	0.97	0.90	0.80	**0.87**	17.162
	LSH-Div	0.89	0.80	0.76	0.81	1.265	0.68	0.88	0.81	0.72	2.392	0.53	0.89	0.80	0.62	4.437
	LSH-SDiv	0.91	0.78	0.76	**0.82**	**0.986**	0.69	0.88	0.80	0.73	2.417	0.52	0.88	0.80	0.61	3.537
MMR [13]	NH	0.92	0.73	0.68	0.77	5.168	0.95	0.86	0.75	0.83	10.585	0.96	0.90	0.76	0.84	16.524
	LSH-Div	0.91	0.77	0.73	0.80	1.135	0.91	0.92	0.85	0.87	2.378	0.87	0.97	0.89	0.88	3.828
	LSH-SDiv	0.92	0.78	0.75	**0.81**	**1.102**	0.93	0.91	0.84	**0.88**	**2.085**	0.89	0.96	0.88	**0.88**	**4.106**
QP-Rel	NH	1.00	0.74	0.69	0.81	704.9	1.00	0.82	0.73	0.84	947.09	1.00	0.87	0.76	0.86	1137.19
	LSH-Div	0.93	0.80	0.77	0.83	0.487	0.92	0.94	0.86	0.88	0.499	0.86	0.98	0.90	0.88	0.502
	LSH-SDiv	0.97	0.78	0.74	**0.83**	**0.447**	0.96	0.89	0.82	**0.88**	**0.464**	0.93	0.95	0.86	**0.89**	**0.473**

MMR [13]. In Table 3.1, we denote NH as Null Hash i.e, without using any hash function, LSH-Div with the hash function in Algorithm 1 and LSH-SDiv with the hash function in Algorithm 2.

We can see in Table 3.1, that our hash functions in combination with various methods are superior to the methods with NH. Our extensions based on LSH-Div and LSH-SDiv hash functions out-perform in all cases with respect to the h-score. Interestingly, LSH-Div and LSH-SDiv with NN report maximum h-score than any other methods. This observation implies that diversity can be preserved in the retrieval by directly using standard LSH based nearest neighbor method. We also report a significant speed up even for a moderate database of $35K$ images. Readers familiar with LSH will also agree that our methods will enjoy better speed up in presence of larger databases and higher dimensional representations.

In Table 3.1 the greedy method with our hash functions reports very low precision at top-20 and top-30 retrievals. This indicates that the greedy method may sometimes pick points too far from the query and might report images that are not relevant to the query. This observation is illustrated with our toy dataset in Figure 3.1. Notice that the existing diverse retrieval methods with NH report diverse images, but they are highly inefficient with respect to the retrieval time. Especially, the QP-Rel method also needs unreasonable memory for storing the gram matrix. To avoid any memory leaks, we partitioned the images into seven (number of categories) blocks and evaluated the queries independently i.e., when the query is flower, we only look at the block of flower images and retrieve diverse set of flowers. Although

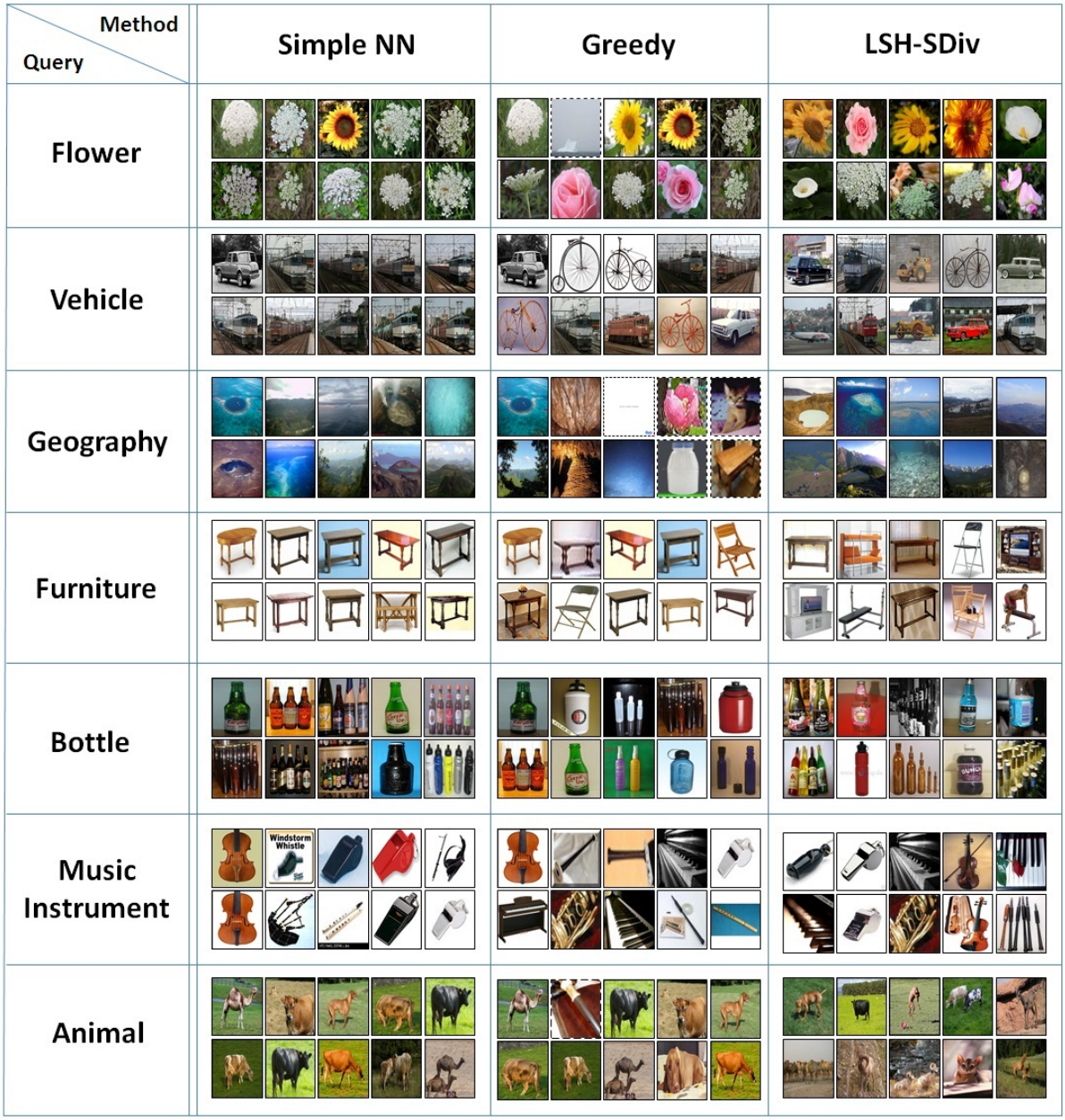


Figure 3.4: In the plot, we show qualitative results for seven example queries from the ImageNet database. Top-10 retrieved images are shown for three methods: the first column with the simple NN method, the second column with Greedy method, and the third column with the proposed LSH-SDiv method. The images marked with dotted box are the incorrectly retrieved images with respect to the query. Notice that the greedy method fails to retrieve accurate retrieval for some of the queries. Our method, consistently retrieves relevant images and simultaneously shows better diversity.

the QP-Rel method achieves better diversity, it is still computationally very expensive. Having such partitions is highly impractical and not feasible for other large datasets. We therefore, omit the results using QP-Rel method on other datasets.

3.5.2 Multi-label Classification

We use one of the largest multi-label datasets, LSHTC, to show the effectiveness of our proposed method. This dataset contains the Wikipedia documents with more than $300K$ labels. To avoid any bias towards the most frequently occurring labels, we selected only the documents which have at least 4 or more labels. Thus, we have a data set of $754K$ documents with $259K$ unique labels. For our experiment, we randomly divide the data in 4:1 ratio for training and testing respectively. We use the large scale multi label learning (LEML) [108] algorithm to train a linear multi-class classifier. This method is shown to provide state of the art results on many large multi label prediction tasks.

In Table 3.2, we report the performance of the label prediction with LEML and compare with our methods that predict diverse labels efficiently. Since, the number of labels for each document varies, we used a threshold parameter to limit the number of predicted labels to the documents. We selected the threshold by cross validating such that it maximizes the h-score. The precision and recall values corresponding to this setting are shown in the table.

In LSHTC3 dataset, the labels are associated with a category hierarchy which is cyclic and unbalanced i.e., both the documents and subcategories are allowed to belong to more than one other category. In such cases, the notion of diversity i.e., the extent to which the predicted labels belong to multiple categories can be estimated using drank [64]. Since, the category hierarchy graph is cyclic, we prune the hierarchy graph to obtain a balanced tree by using breadth first search (BFS) traversal. The diversity of the predicted labels is computed as the drank score on this balanced tree. In Table 3.2, we report the overall performance of a method in terms of h-score i.e., the precision and the drank score.

As can be seen from Table 3.2, the LSH-Div method shows a reasonable speed-up but fails to report many of the accurate labels i.e., has low precision. Since, the LSHTC3 dataset is highly sparse in a large dimension, random projections generated by LSH-Div method are a bit inaccurate and might have resulted in poor accuracy. The proposed LSH-SDiv approach significantly boosts the accuracy, since, the random vectors in the hash function are projected onto the principal components that capture the data distribution accurately. The results shown in table are obtained by using 100 random projections for both LSH-Div and LSH-SDiv hash functions. For the LSH-SDiv method, we project the random projections onto the top 200 singular vectors obtained from the data points. Clearly, LSH-SDiv based hash function improves the diversity within the labels and outperforms LEML, MMR, LSH-Div, and PCA-Hash methods in terms of overall performance (h-score). We also obtain a speed-up greater than 20 over LEML method and greater than 80 over MMR method on this dataset. Note that, we omitted the results with greedy method as they failed to report accurate labels on this large dataset.

Table 3.2: Results on LSHTC3 challenge dataset with LEML, MMR, LSH-Div, PCA-Hash and LSH-SDiv methods. LSH-SDiv method significantly outperforms both LEML, MMR, LSH-Div and PCA-Hash methods in terms of overall performance, h-score as well as the retrieval time.

Method	P	R	f-score	D	h	time (msec)
LEML [108]	0.304	0.196	0.192	0.827	0.534	137.1
MMR [13]	0.275	0.134	0.175	0.865	0.418	458.8
LSH-Div	0.144	0.088	0.083	0.825	0.437	7.2
PCA-Hash	0.265	0.096	0.121	0.872	0.669	5.9
LSH-SDiv	0.318	0.102	0.133	**0.919**	**0.734**	**5.7**

Table 3.3: Results on Tag Suggestion for Flickr Images. Notice that LSH-SDiv method improves the diversity even when the accuracy is low. Note that, we omitted the results with other diverse retrieval approach since, they fail to report accurate tags on this dataset.

Method	P@1	P@3	P@5	D	h	time (msec)
NN [56]	0.057	0.054	0.053	0.910	0.100	472.1
LSH-Div	0.048	0.037	0.034	0.911	0.065	**3.0**
LSH-SDiv	0.051	0.047	0.039	**0.915**	0.076	4.6

3.5.3 Image Tagging

In the task of image tagging, our goal is to predict/assign multiple tags (text labels) to a query image. It was shown [56] that the relevance of a tag with respect to an image might be inferred from tagging behavior of visual neighbors of that image. Essentially, the common tags are propagated through visual links introduced by visual similarity and then, the relevance of each tag is computed using a neighbor voting scheme. For a given query image, we first find the top-k relevant neighbors from the annotated images and score the tags by voting received from these images. We rank the tags by using the tag relevance score as described in [56].

For our experiments, we use a large collection of annotated Flickr images from [56]. For each image, we extract a combined 64-dimensional global feature and use it to compute the visual similarity between the pair of images. By removing images which failed to extract visual features, we obtained 2.7 Million labeled images which has 5,09,234 unique tags, with an average value of 5.4 tags per image. We illustrate the performance of tag suggestion for unlabelled images on a test set of 314 Flickr images as used in [56]. We fix the number of visual neighbors to 500 as suggested in [56] for the visual neighbor search. For the LSH-Div and LSH-SDiv methods, we use all the visual neighbors retrieved by the respective methods. Considering the evaluation scheme adopted in [56], for each method, we select top 5 tags as the final suggestions for each test image. Diversity between a pair of tags is computed using a Wordnet based semantic similarity measure.

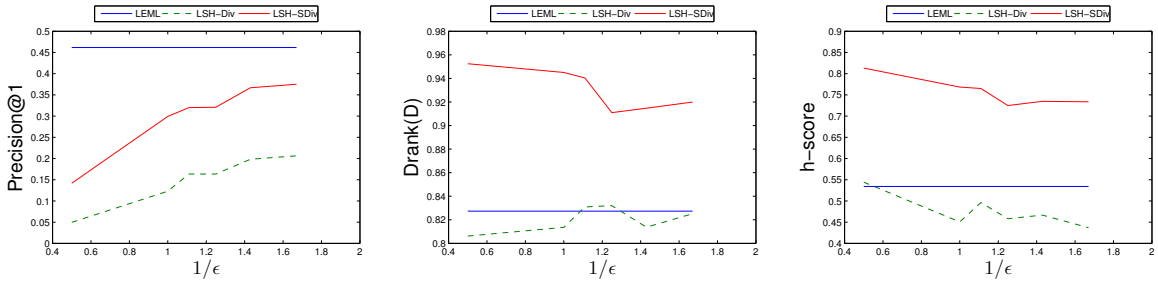
Query Image / Method						
Nearest Neighbor	Cloud, sky, mountain, blue, water	Flower, red, macro105mm, rose, nature	Light, fire, night, camp, sunset	Race, bike, partial, bicycle accident	Flower, red, macro105mm, pink, garden	Tree, hike, park, house, mountain
LSH-Div	Sky, snow, snowboard, winter, italia	Chartact, red, car, canada, grape hyacinth	---	Adult and juvenil, life, leavalley, restal, kiss	Flower, orangad, macro105mm, red, rose	---
LSH-SDiv	Mountain, travel, sky, cloud, lake adelaid	Flower, red, rose, garden, lea valley	Light, night, firework, flower, concert danzig	Sanfrancisco, bike rack, bike, tour of california	Flower, macro105mm, red, nature, green	Tree, fall, autumn, car

Figure 3.5: Top-5 tags obtained by different methods on Image Tagging Task. First row is the query image. Subsequent rows show tags predicted by simple NN method, the LSH-Div and the LSH-SDiv methods. Observe that LSH-SDiv method retrieves better tags than the LSH-Div method, respectively. Here, (- -) indicates that all the tags have received equal votes.

In Table 3.3, we report the performance of all the three methods in terms of precision at 1, 3 and 5. We see that the diversity among the tags is improved with the LSH-Div and LSH-SDiv methods. Notice that the overall performance in terms of h-score is low for LSH based methods. This is because all the three methods have very poor precision. Note that, we omitted the results with other diverse retrieval methods as they failed to report accurate tags on this dataset. Nevertheless, our methods show an impressive speed-up of 100 over the exhaustive NN on this dataset.

3.6 Discussions

Our empirical evidences from the three different experiments confirm that we have a high precision for the image category retrieval scenario, a medium precision for the multi-label classification, and low precision for the image tagging scenario. The proposed algorithm is effective and robust, since, it improves diversity for different levels of accuracies, i.e., low, medium and high. In particular, our algorithm performs especially better when the accuracy is medium or high. Obviously, all methods show equally diverse solution when no relevant images/labels exist in the neighbor set. Our approach comes with an additional advantage of being more efficient computationally, which is crucial for large datasets.



(a) (b) (c)

Figure 3.6: Results on LSHTC3: LSH-Div and LSH-SDiv methods use 100 hash functions. (a) LSH-SDiv method gives better precision than LSH-Div method. (b) LSH-SDiv method shows better diversity than LEML and LSH-Div methods. (c) LSH-SDiv method performs significantly better than the LEML and LSH-Div methods in terms of h-score.

Figure 3.6 illustrates the performance on LSHTC3 dataset with respect to the parameter ϵ. In the figure we show the performance obtained when 100 random projections are selected for the LSH-Div method. For the LSH-SDiv method we project the 100 random projections onto the top-200 singular vectors obtained from the data. Notice that the conventional LSH hash function considers only random projections and fails to five good accuracy. As discussed in Section 3.1.2, a large number of random projections are needed to retrieve accurate labels, which would slow down the retrieval procedure.

In contrast, the LSH-SDiv method can successfully preserve the distances i.e., report accurate labels by projecting onto a set of β principal components if the data is embedded in β dimensions only. Similarly, if the $\beta + 1$-th singular value of the data matrix is $\sigma_{\beta+1}$ then the distances are preserved up to that error and has no dependence on say ϵ that is required by standard LSH hash function. Hence, LSH-SDiv based technique typically requires much smaller number of hash functions than the standard LSH method and hence, is much faster as well (see Table 3.1 and Table 3.2).

In conclusion, the proposed LSH-SDiv method significantly outperforms the baselines and the standard LSH-Div method in terms of accuracy, diversity and retrieval time. Moreover, the LSH-SDiv method achieves this using very less number of hash-bits. Although locality sensitive hashing methods are strongly motivated for solving the approximate nearest neighbor search efficiently, we showed its ability to retrieve diverse with competitive performances.

Chapter 4

Diversity using Learned Metrics

> *Multimedia databases (in particular image databases) are different from traditional system since they cannot ignore the perceptual substratum on which the data come. There are several consequences of this fact. The most relevant for our purposes is that it is no longer possible to identify a well defined meaning of an image and, therefore, matching based on meaning is impossible. Matching should be replaced by similarity assessment and, in particular, by something close to human preattentive similarity* - Similarity is a geometer [83]

4.1 Instance based Image Retrieval

Instance-level image retrieval algorithms have gained recent prominence because of their applicability to two main areas, image recognition for product search like retail products [30, 87] and localization [58, 86]. The task of searching in an image database for specific "instances" of an object or subject is called instance retrieval (IR). For example, when searching for images of "Maruti car", a generic image search might retrieve various cars like Maruti, Toyata, BMW, etc. IR algorithms, on the other hand, retrieve images of various models like "Maruti Celerio", "Maruti Swift" etc. IR methods are expected to perform under several physical constraints like variations in viewpoint of camera, time of day, camera zoom etc.

The success of IR algorithms usually depends on the low-level image features, such as color, texture, and shape, that represent the visual content present in the images. The most popular image representation is the Bag of Visual Words (BoVW) model [67]. A typical BoVW pipeline for representing images is composed of the following steps: (i) **extracting** the local features from each image, (ii) **encoding** the local features to the corresponding visual words and (iii) performing **spatial binning**. Initially, a large set of local features are extracted from a training image corpus. These features are clustered to divide the local feature space into informative regions (called "visual words") and the collection of the obtained visual words is the visual vocabulary. Feature extraction is carried using the popular SIFT [60] which is designed to capture appearance and local image structures that are invariant to image transformations such as translation, rotation, and scaling. Next, in the encoding step, the local features of an image are

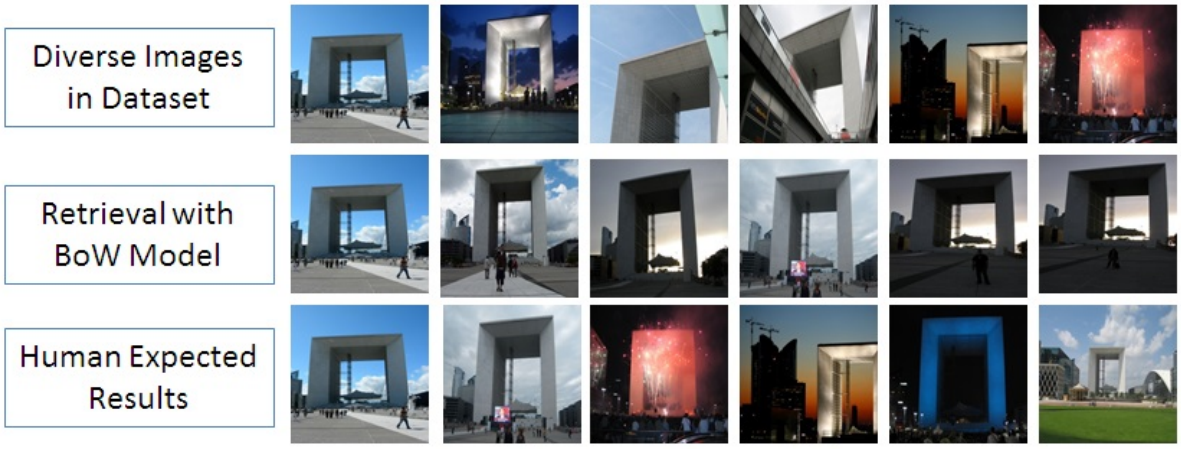


Figure 4.1: Sample images of the Paris dataset for the monument "La Grande Arche de la Dfense". Top row shows the diverse images present in the dataset. Middle row shows a sample query image, and corresponding retrieved results using a BoVW model. Bottom row shows the human expected diversity in results. (*Images best viewed in color*)

assigned to the nearest visual word's centriod (in Euclidean distance) and a histogram of visual words is generated. Finally, spatial information is encoded by dividing the image into several (spatial) regions, compute the encoding of each region and concatenating all the resulting histograms. Thus, in IR, when a query image is given, one computes the SIFT features and encodes the visual information in the form of a histogram and retrieves relevant images that are close in the Euclidean distance.

However, when a database has many similar images, IR methods result in near identical images at the top. This is because of two properties: Firstly, local features like SIFT are more adept at identifying near identical images and often confuse between different views of the same image and different but similar looking images. Such a differentiation requires higher order features to be computed. Secondly, these approaches do not penalize duplicate results aggressively. Therefore, the retrieved results are more homogeneous with little diversity. We define diversity in IR as accurate retrieval of instances that show variations in physical properties like geometry and illumination.

Further more, in the BoVW methods the distance metric, often pre-defined, used for image similarity is detrimental to accuracy and diversity of the results. This limits the capacity of IR algorithms, because they usually assume that the distance between two similar objects is smaller than the distance between two dissimilar objects. This assumption may not hold, especially in the case of IR when the input space is heterogeneous i.e., diverse in visual content. For instance, the outdoor images (like monuments) are most effected by natural light, position from the which images are captured, and camera zoom that is intrinsic property of an image. Product search might have other properties like occlusion, but this is out of scope of this work.

We illustrate these characteristics with an example in Figure 4.1. Given a dataset containing several distinct views of a monument (La Grande Arche de la Defense in Paris, first row, Figure 4.1), BoVW based algorithms [67, 71] typically retrieve near similar results for a query image (second row,

Figure 4.1), even when the database itself contains diverse images. It can be easily seen that users searching for images of La Grande Arche de la Defense, might better appreciate the set of results shown in the third row of Figure 4.1, because its diversity in viewpoint, camera zoom, time of day etc. gives much better visual understanding of the monument itself. We thus make the case that diversity is an important characteristic for an IR algorithm to have.

Metric Learning: We show that the key to encoding diversity is to find appropriate distance metric which allows for variations these physical properties. Learning distance metric from available domain has attracted much interest in recent studies [9, 55, 59]. The domain information is usually cast in the form of two pairwise constraints: must-link and cannot-link constraints. The must-link constraints enforce smaller distances for the pair of "similar" objects, and cannot-link constraints enforce large distances for the pair of "dissimilar" objects. The optimal distance metric is found such that majority of these pairwise constraints are satisfied. Our goal is to learn this distance metric (A), under certain physical constraints, to improve diversity in IR.

We use a popular metric learning approach called Information theoretic metric learning (ITML) [23]. ITML algorithm uses an information-theoretic cost model which iteratively enforces pairwise similarity/dissimilarity constraints, yielding a learned Mahalanobis distance metric, A. The Mahalanobis distance is a bijection to a Gaussian distribution with its covariance set as an inverse of A. Exploiting this bijective property, ITML poses the metric learning problem as a convex optimization of a relative entropy between a pair of Gaussian distributions with unknown A and the identity I or A_0 a prior knowledge about the inter-point distances, under simple distance similar(S)/dissimilar(D) constraints.

$$\begin{aligned} \min \;& D_{ld}(A, A_0) \\ \text{s.t.} \;& A \succeq 0 \\ & d_A(x_i, x_j) \leq u \qquad (i,j) \in S \\ & d_A(x_i, x_j) \geq v \qquad (i,j) \in D \end{aligned} \tag{4.1}$$

where, $D_{ld}(A, A_0) = tr(AA_0^{-1}) - \log det(AA_0^{-1}) - d$; v and u are large and small values, respectively. Solving Eq.(4.1) involves repeatedly projecting the current solution onto a single constraint, via an update:

$$A_{t+1} = A_t + \beta_t A_t (x_{i_t} - x_{j_t})(x_{i_t} - x_{j_t})^T A_t, \tag{4.2}$$

In the equation, x_{i_t} and x_{j_t} are the constrained data points for iteration t, and β_t is a projection parameter computed by the ITML algorithm. This formulation regularizes the optimization problem so as to seek a metric that satisfies the given constraints and is closest to the Euclidean distance.

To summarize, our algorithm executes in three phases: i) perform metric learning using ITML, Algorithm 4 in line (1), to find appropriate metric and ii) transform the initial feature space to $\mathcal{T}$ using $A^{\frac{1}{2}}$, Algorithm 4 in line (2), and iii) find the k nearest neighbors, Algorithm 4 in lines (4-7), to report the set of retrieved images, $R_q \in \mathcal{X}$. Throughout the algorithm, several variables are used that are specific to the quality of diversity. The essential control variables that direct the behaviour of the algorithm are:

Algorithm 4: Diverse Retrieval Using Metric Learning

Input: $\mathcal{X} = \{x_1 \ldots, x_n\}$, where $x_i \in R^d$, a query $q \in \mathbb{R}^d$ and k an integer. $A_0 = I$, is prior about the inter-point distances. S, D are similar and dissimilar constraints.

$A \leftarrow$ ITML($\mathcal{X}$, A_0, S, D u, v) *// Learn metric*

$\mathcal{X} \rightarrow \mathcal{T} = A^{\frac{1}{2}}\mathcal{X}$ *// Apply linear transformation*

$\mathcal{R}_q \leftarrow \phi$ *// Retrieved images*

for $i \leftarrow 1$ ***to*** k **do**

  $t^* \leftarrow \text{argmin}_{(t \in \mathcal{T})}(\|q - t\|^2)$

  $\mathcal{T} \leftarrow \mathcal{T} \setminus t^*$

  $\mathcal{R}_q \leftarrow \mathcal{R}_q \cup t^*$

Output: $\mathcal{R}_q$, set of retrieved images

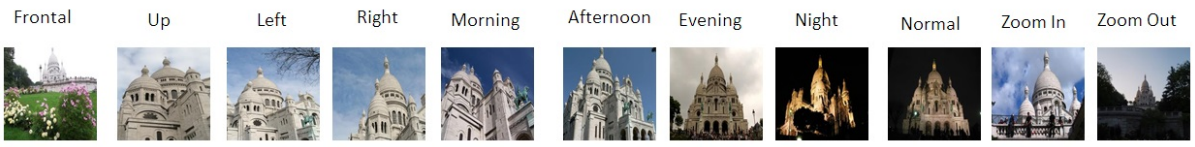


Figure 4.2: Images with labels for the "Sacre Coeur" monument in the Paris dataset. (*Images best viewed in color*)

i) the choice of u and v in the ITML and ii) the number of constraints, $|S| + |D|$. See Algorithm 4 for a detailed description of our approach.

4.2 Exploring Diversity from Semantics

Metric learning can be seen as a data-driven transferring of semantic information from the class labels to input feature space. In order to learn a appropriate metric for BoVW features to promote diversity in the retrieval, we have to define the constraints $d_A(x_i, x_j) \leq u$ or $d_A(x_i, x_j) \geq v$ for a pair of feature vectors x_i and x_j, corresponding to images that are similar and dissimilar, respectively. In this work, physical properties are assigned as class labels (refer Figure 4.2) while BoVW forms the feature space.

In order to perform an IR, we first extract SIFT [60] features from the input query image and compute visual words using the cluster centers of the database to be searched. In our experiments, we extract 100 visual words using the popular VLFeat library [94]. We set the variables v and u in Eq.(4.1) to the 97^{th} and 3^{rd} percentiles of the distribution of pairwise Euclidean distances within the dataset, respectively. We randomly sample 100 pairwise constraints from a pool of annotated images to learn the distance metric and, apply the transformation on the basic BoVW features as discussed in Section 4.1. As a result, the matching procedure using the learned distance metric takes the same time as the BoVW method.

Table 4.1: Paris Dataset: Comparison of the retrieval performance for BoVW and learned metrics using top-5 results. V- Viewpoint, T - Time of Day, Z- Zoom, Div-Diversity. H is the harmonic mean of accuracy with their respective diversity scores. Notice the best performances are marked in **bold**.

Method	Accuracy	V-Div	H-V	T-Div	H-T	Z-Div	H-Z
BoVW [67]	0.817	**0.412**	**0.511**	0.537	0.625	0.384	0.445
ITML [23]	**0.822**	0.391	0.495	**0.592**	**0.652**	**0.434**	**0.474**

Table 4.2: User Study: Results averaged over 210 queries, answering which method produced more useful results.

Method	BoW	Our Approach	Tie
User Preference	84/210 = 40%	**97/210 = 46.19%**	29/210 = 13.81%

1) Datasets: We use the Paris dataset [72] which consists of approximately 6K high quality (1024 X 768) images of monuments in Paris like La Defense and Pantheon. Note that this collection of Paris images is considered to be a challenging dataset. Since the images are not tagged based on monument visibility, we manually annotated 200 images with 12 labels in the following categories: viewpoint (frontal, up, down, left, right), camera zoom (zoomed in, zoomed out, normal), time of day (morning, afternoon, evening and night). Figure 4.2 shows labels for a sample monument image in the Paris dataset.

2) Evaluation Criteria: There is no evaluation metric that seems to be universally accepted as the best for measuring the performance of methods that aim to obtain diverse retrieval [75]. Diversity necessarily depends on the collection over which the search is being run [35, 36]. Diversity also depends on a system's performance at basic ad hoc retrieval i.e., how many images are relevant to the user query. Therefore, similar to precision and recall, there is a need to balance between accuracy and diversity in the retrieval. In this work, we keep a balance between accuracy and diversity by maximizing the harmonic mean of these two criteria. Below, we describe the performance measures used to evaluate our experiments.

Accuracy: We measure the accuracy of the retrieval in terms of the proportion of relevant images (to the given query) in the retrieved results, aggregated over 50 trails.

Diversity: We measure diversity in terms of the entropy as $-\Sigma_{i=1}^{m} s_i \log s_i$, where s_i is the fraction of images of i^{th} tag, and m is the number of possible labels, aggregated over 50 trails.

Empirical Results: To empirically evaluate the methods, we pick 50 random query images and retrieve results for these queries from the 200 labeled images. We use the labels of the top-5 results to compute accuracy and diversity scores. As discussed above, we show in Table 4.1 the overall perfor-

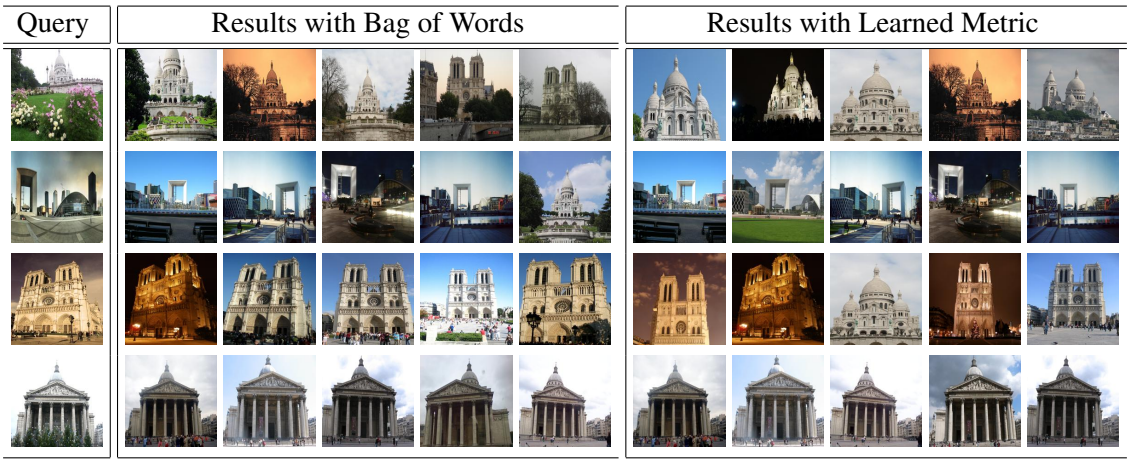


Table 4.3: Four pairs of retrieval results from the Paris dataset. Top-5 candidates are shown for visual comparison between BoVW and Learned metric based approaches. For each query (column one from top to bottom), accuracy for BoVW and ITML Methods are $\{0.6, 0.8, 1, 1\}$ and $\{1, 1, 0.8, 1\}$, respectively. (*Images best viewed in color*)

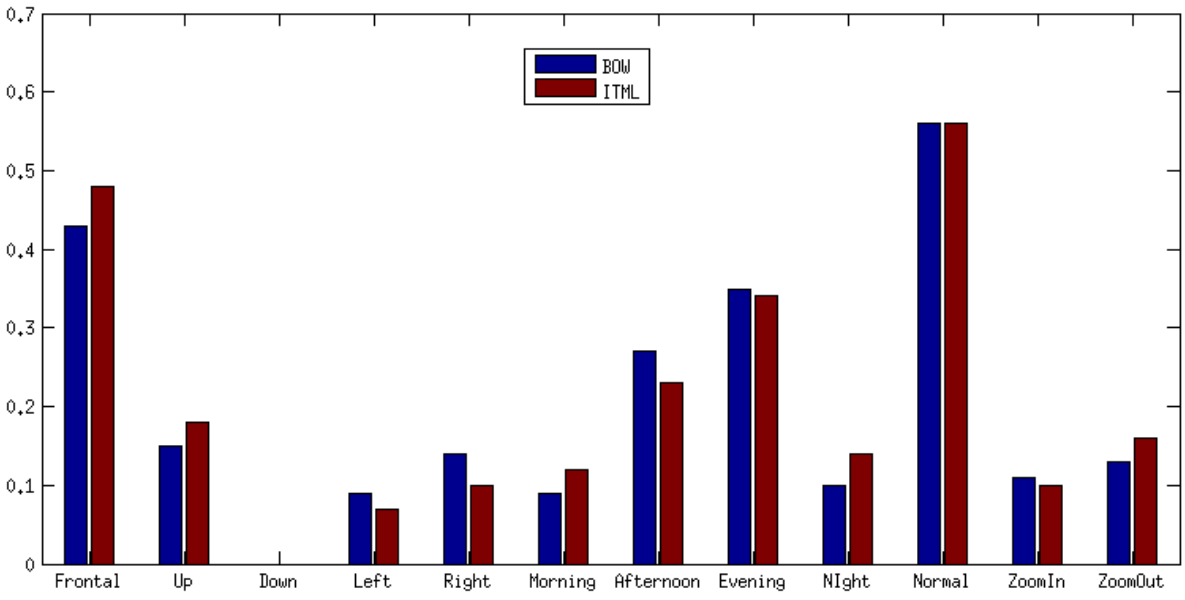


Figure 4.3: Histogram of labels over 50 queries for BoVW and ITML based retrieval algorithms. Notice the improvements in the "Morning", "Night" using ITML approach and also note the rise in "Zoom Out" label with a drop in "ZoomIn" label. (*Image best viewed in color*)

mance measure as the harmonic mean of accuracy and diversity. We report results for viewpoint, time of day and camera zoom diversity.

In our results, we observe an improvement in the accuracy of the retrieval using ITML. Notice that ITML outperforms BoVW model in terms of h-score. This demonstrates the effectiveness of using metric learning to obtain both relevant and diverse set of monument images. In order to measure diversity, we use the distribution of the histogram labels (in Figure 4.3), with an equal distribution over all labels being the most desirable result. Notice how ITML improves the "night" and "morning" labels by suppressing the "afternoon" and "evening" labels. We also see a rise in "ZoomOut" label with a drop in "ZoomIn" label. It is important to notice that the diversity with respect to viewpoint is low for ITML approach (see Table 4.1 column 2), and this can also be observed in the rise of spikes at "Frontal" and "Up" labels for ITML approach in Figure 4.3.

The first rows of Table 4.3 give a visual representation of the top 5 retrieved images given the query images (shown in the first column). Note how the retrieved results are visually very similar to the query image in many aspects like appearance, viewpoint, zoom and even to some extent the time of day. This highlights the problem that we alluded to earlier, about the absence of diversity in results with traditional BoVW model. As can be seen in Table 4.3, our approach shows a greater visual diversity in the retrieved images. These visual results convincingly prove the ability of learning metrics (from pairwise constraints) can be helpful to improve the diversity by as much as 5% (in the case of time of day and camera zoom, refer Table 4.1 column 5) while still retaining similarity among the results.

Evaluating Human Expectations: We evaluate the utility of our approach based on testimonials from 14 different users randomly selected for trails. We asked them to rate 5 queries by pointing out which among IR method between BoVW and our approach gave the most relevant results. We the averaged results for the 210 queries i.e., 14 users X 5 images X 3 criteria. Table 4.2 shows that users in general rated our approach superior to the BoVW based IR approach.

In conclusion, we proposed a metric learning-based diverse IR method and presented a systematic experimental comparison with traditional bag of visual words model. Although retrieving visually similar images is arguably the most obvious application where metric distance learning plays an important role, we showed its application to diverse IR where a good distance metric is essential for obtaining competitive performances.

Chapter 5

Conclusions and Outlook

Diversity in Retrieval: Randomize don't optimize

Diversity is a desirable property of an information retrieval system that we should seek when query intent is uncertain. This is a challenging task, perhaps in part because, there is a wide diversity in what diversity means. Conventional methods make certain assumptions on the choice of the distance functions and accordingly try to solve the appropriate problem at hand. Moreover, existing approaches are inefficient and do not scale to large databases. This inherently limits the scope and utility of the retrieval systems that we can efficiently deploy today. Therefore, the "sole objective" of this thesis is to design efficient algorithms that should be adaptable to any target domain.

In this thesis, we focus on the retrieval task that is formulated as the problem of finding nearest neighbors to the user query. We observe that, in most of the cases, one needs to trade-off accuracy for diversity. That is, rather than finding the nearest neighbor, we would need to select a point which is a bit farther from the given query but is dissimilar to the other retrieved points. Hence, we would need to find approximate nearest neighbors while ensuring that the retrieved points are diverse.

Following this intuition, this thesis has proposed methods to address the following key challenges to design diverse retrieval methods: a) accurate, b) diversity, and c) sub-linear retrieval time. We present an approach to efficiently retrieve diverse results based on locality sensitive hashing. A careful application of randomness provides an elegant solution to retrieve relevant and diverse results in sub-linear retrieval time.

While the approaches proposed herein are by no means the complete and final solutions to the diverse retrieval problem, they represent real progress towards designing effective diverse retrieval algorithms. A particularly salient feature in the approach presented herein is that they address these challenges by identifying a fundamental issue of practical importance, and thus motivate models which directly try to tackle deep research questions like a) using randomness to find optimal solutions to diverse retrieval problem, and b) learning metrics which bridges the gap between the high-level semantic information and the low-level visual information in the images to encourage diversity in the image retrieval.

The contributions of this thesis include (1) methods for developing hash functions that can retrieve diverse results for user queries, (2) extension of our method to diverse multi-label prediction, (3) demonstrate the possibility of retrieving accurate and diverse results in sub-linear time, and (4) demonstrate the possiblilty of learning metrics from high-level semantic information to retrieve diverse images.

For three different applications, our proposed methods retrieve significantly more diverse and accurate data points, when compared to the existing methods. The results obtained by our approach are appealing: a good balance between accuracy and diversity is obtained by using only a small number of hash functions. We obtain $100x$-speed-up over existing diverse retrieval methods while ensuring high diversity in retrieval.

Further, we also highlight the need for learning metrics to captures different aspects of images characteristics to improve diversity in image retrieval. For the instance based image retrieval, we demonstrate that the traditional BOW model retrieve results that are visually very similar to the query image in many aspects like appearance, viewpoint, zoom and even to some extent the time of day. We have proposed a simple method for efficiently retrieving diversely similar results for a given query image. We have shown the efficacy and performance of our approach with extensive experiments using a real dataset.

Future Perspectives

In this final section, we discuss how related research fields can also benefit from the methods proposed in this dissertation, as well as more general information retrieval problems.

We believe that other approximate nearest neighbor retrieval algorithms like Randomized KD-Trees also encourage diversity in the retrieval. In our case, the rigorous theory of locality sensitive hashing functions naturally supports its performance in relevance, diversity and retrieval time. Note that the random hash functions designed in our methods are only geared to maintain spread among points with very high probability. While doing so, the algorithm has no way of knowing which solutions are diverse and which are not diverse. Therefore, from the optimization perspective, for these methods the task of providing any guarantees of the true solution to the diverse retrieval problem is challenging. With this respect, it would be interesting to examine the existence of approximation guarantees to the optimal solution.

From the learning perspective, applicability of our approach to many other retrieval scenarios involving different definitions of diversity (visual, temporal, spatial, and topical aspects), knowledge source combination (image and text), interactive retrieval systems (relevance feedback), and so forth are of immediately useful extensions. This requires us to move away from the approaches that are ignorant of user context, and towards methods that can learn rich, structured models from feedback collected via user interaction. This line of research can lead to highly efficient information systems where the prior information of the retrieval task is available immediately.

Related Publications

The work in my thesis has been disseminated to the following conferences.

- Vidyadhar Rao, Prateek Jain, and C. V. Jawahar. "**Diverse Yet Efficient Retrieval using Hash Functions**", in *arXiv preprint*, `http://arxiv.org/abs/1509.06553`, Sep 22, 2015.
- Vidyadhar Rao, Ajitesh Gupta, Visesh Chari, and C.V. Jawahar. "**Learning Metrics for Diversity in Image Retrieval**", in *The Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, NCVPRIPG, 2015.
- Vidyadhar Rao, and C. V. Jawahar. "**Semi-supervised Clustering by Selecting Informative Constraints**", in 5^{th} *International Conference on Pattern Recognition and Machine Intelligence*, PReMI, 2013.

Bibliography

[1] R. Agrawal, A. Gupta, Y. Prabhu, and M. Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *WWW*, 2013.

[2] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 2000.

[3] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*. IEEE, 2006.

[4] T. Arni, J. Tang, M. Sanderson, and P. Clough. Creating a test collection to evaluate diversity in image retrieval. *Beyond binary relevance: preferences, diversity and set-level judgments*, 2008.

[5] H. Avron, P. Maymounkov, and S. Toledo. Blendenpik: Supercharging lapack's least-squares solver. *SIAM Journal on Scientific Computing*, 2010.

[6] R. Basri, T. Hassner, and L. Zelnik-Manor. Approximate nearest subspace search. *TPAMI*, 2011.

[7] S. Basu, A. Banerjee, and R. Mooney. Semi-supervised clustering by seeding. In *Machine learning-international workshop then conference-*, 2002.

[8] S. Basu, A. Banerjee, and R. J. Mooney. Active semi-supervision for pairwise constrained clustering. In *SDM*, 2004.

[9] M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *ICML*. ACM, 2004.

[10] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001.

[11] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*. 2004.

[12] L. Breiman. Random forests. *Machine learning*, 2001.

[13] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.

[14] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *TPAMI*, 2007.

[15] B. Carterette. An analysis of np-completeness in novelty and diversity ranking. *Information Retrieval*, 2011.

[16] L. Cayton and S. Dasgupta. A learning framework for nearest neighbor search. In *NIPS*, 2008.

[17] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC*, 2002.

[18] I. H. Consortium et al. A haplotype map of the human genome. *Nature*, 2005.

[19] W. S. Cooper. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 1973.

[20] W. S. Cooper. On selecting a measure of retrieval effectiveness part ii. implementation of the philosophy. *Journal of the American Society for Information Science*, 1973.

[21] C. K. Dagli, S. Rajaram, and T. S. Huang. Utilizing information theoretic diversity for svm active learn. In *ICPR*, 2006.

[22] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*. ACM, 2004.

[23] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*. ACM, 2007.

[24] F. Delbos and J. C. Gilbert. Global linear convergence of an augmented lagrangian algorithm for solving convex quadratic optimization problems. 2003.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[26] T. Deselaers, T. Gass, P. Dreuw, and H. Ney. Jointly optimising relevance and diversity in image retrieval. In *CIVR*, 2009.

[27] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 2008.

[28] G. Ference, W.-C. Lee, H.-J. Jung, and D.-N. Yang. Spatial search for k diverse-near neighbors. In *CIKM*, 2013.

[29] N. Fuhr. A probability ranking principle for interactive information retrieval. *Information Retrieval*, 2008.

[30] M. George and C. Floerkemeier. Recognizing products: A per-exemplar multi-label image classification approach. In *ECCV*. 2014.

[31] B. Georgescu, I. Shimshoni, and P. Meer. Mean shift based clustering in high dimensions: A texture classification example. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003.

[32] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, et al. The international hapmap project. *Nature*, 2003.

[33] J. Gillenwater, A. Kulesza, and B. Taskar. Near-optimal map inference for determinantal point processes. In *NIPS*, 2012.

[34] A. Gionis, P. Indyk, and R. Motwani. Similarity Search in High Dimensions via Hashing. In *VLDB*, 1999.

[35] P. B. Golbus, J. A. Aslam, and C. L. Clarke. Increasing evaluation sensitivity to diversity. *Information Retrieval*, 2013.

[36] P. B. Golbus, V. Pavlu, and J. A. Aslam. What we talk about when we talk about diversity.

[37] K. Grauman and T. Darrell. Pyramid match hashing: Sub-linear time indexing over partial correspondences. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007.

[38] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 2011.

[39] J. He, H. Tong, Q. Mei, and B. Szymanski. Gender: A generic diversified ranking algorithm. In *NIPS*, 2012.

[40] T. K. Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1998.

[41] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *ACM symposium on Theory of computing*, 1998.

[42] P. Jain, B. Kulis, and K. Grauman. Fast image search for learned metrics. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008.

[43] P. Jain, S. Vijayanarasimhan, and K. Gorauman. Hashing hyperplane queries to near points with applications to large-scale active learning. In *NIPS*, 2010.

[44] N. Jammalamadaka, V. Pudi, and C. Jawahar. Efficient search with changing similarity measures on large multimedia datasets. In *Advances in Multimedia Modeling*. 2006.

[45] J. Jancsary, S. Nowozin, and C. Rother. Learning convex qp relaxations for structured prediction. In *ICML*, 2013.

[46] W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 1984.

[47] S. Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*. IEEE, 1998.

[48] H. A. Khan, M. Drosou, and M. A. Sharaf. Dos: an efficient scheme for the diversification of multiple search results. In *SSDBM*, 2013.

[49] M. K. Kozlov, S. P. Tarasov, and L. G. Khachiyan. The polynomial solvability of convex quadratic programming. *USSR Computational Mathematics and Mathematical Physics*, 1980.

[50] A. Krause and D. Golovin. Submodular function maximization. *Tractability: Practical Approaches to Hard Problems*, 2012.

[51] O. Kucuktunc and H. Ferhatosmanoglu. λ-diverse nearest neighbors browsing for multidimensional data. *TKDE*, 2013.

[52] A. Kulesza and B. Taskar. k-dpps: Fixed-size determinantal point processes. In *ICML*, 2011.

[53] B. Kulis and T. Darrell. Learning to hash with binary reconstructive embeddings. In *NIPS*, 2009.

[54] R. M. Larsen. Lanczos bidiagonalization with partial reorthogonalization. *DAIMI Report Series*, 1998.

[55] J.-E. Lee, R. Jin, and A. K. Jain. Rank-based distance metric learning: An application to image retrieval. In *CVPR*. IEEE, 2008.

[56] X. Li, C. G. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE MultiMedia*, 2009.

[57] H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011.

[58] T.-Y. Lin, S. Belongie, and J. Hays. Cross-view image geolocalization. In *CVPR*. IEEE, 2013.

[59] A. López-Méndez, J. Gall, J. R. Casas, and L. J. Van Gool. Metric learning from poses for temporal clustering of human motion. In *BMVC*, 2012.

[60] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[61] G. Lugosi. Concentration-of-measure inequalities. 2004.

[62] M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 2011.

[63] P.-G. Martinsson. Rapid factorization of structured matrices via randomized sampling. *arXiv preprint arXiv:0806.2339*, 2008.

[64] P. K. R. Mittapally Kumara Swamy and S. Srivastava. Extracting diverse patterns with unbalanced concept hierarchy. In *PAKDD*, 2014.

[65] J. Moraleda. Gregory shakhnarovich, trevor darrell and piotr indyk: Nearest-neighbors methods in learning and vision. theory and practice. *Pattern Analysis and Applications*, 2008.

[66] K. G. Murty and F.-T. Yu. *Linear complementarity, linear and nonlinear programming*. Citeseer.

[67] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2161–2168. IEEE, 2006.

[68] E. Ntoutsi, K. Stefanidis, K. Rausch, and H.-P. Kriegel. Strength lies in differences: Diversifying friends for recommendations through subspace clustering. In *CIKM*, 2014.

[69] N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS genetics*, 2006.

[70] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet:: Similarity: measuring the relatedness of concepts. In *HLT-NAACL*, 2004.

[71] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[72] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[73] Y. Prabhu and M. Varma. Fastxml: a fast, accurate and stable tree-classifier for extreme multi-label learning. In *KDD*, 2014.

[74] J. R. Quinlan. Induction of decision trees. *Machine learning*, 1986.

[75] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims. Redundancy, diversity and interdependent document relevance. In *ACM SIGIR Forum*. ACM, 2009.

[76] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *ICML*, 2008.

[77] M. Rastegari, C. Fang, and L. Torresani. Scalable object-class retrieval with approximate and top-k ranking. In *ICCV*, 2011.

[78] P. Ravikumar and J. Lafferty. Quadratic programming relaxations for metric labeling and markov random field map estimation. In *ICML*, 2006.

[79] V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 2009.

[80] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *Computer vision and image understanding*, 2001.

[81] M. Sanderson, J. Tang, T. Arni, and P. Clough. What else is there? search diversity examined. In *Advances in Information Retrieval*. Springer, 2009.

[82] S. Sanner, S. Guo, T. Graepel, S. Kharazmi, and S. Karimi. Diverse retrieval via greedy optimization of expected 1-call@ k in a latent subtopic relevance model. In *CIKM*, 2011.

[83] S. Santini and R. Jain. Similarity is a geometer. *Multimedia Tools and Applications*, 1997.

[84] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003.

[85] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 2011.

[86] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In *CVPR*. IEEE, 2012.

[87] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Mobile product image search by automatic query object extraction. In *ECCV*. 2012.

[88] N. Silver. *The signal and the noise: Why so many predictions fail-but some don't*. Penguin, 2012.

[89] K. Singh and M. Xie. Bootstrap: a statistical method.

[90] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2000.

[91] N. Sundaram, A. Turmukhametova, N. Satish, T. Mostak, P. Indyk, S. Madden, and P. Dubey. Streaming similarity search over one billion tweets using parallel locality-sensitive hashing. *VLDB*, 2013.

[92] P. Tandon. *Learning in Large Scale Image Retrieval Systems*. PhD thesis, International Institute of Information Technology Hyderabad, India, 2009.

[93] P. Tandon, P. Nigam, V. Pudi, and C. V. Jawahar. Fish: a practical system for fast interactive image search in huge databases. In *Proceedings of international conference on Content-based image and video retrieval*. ACM, 2008.

[94] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *ACM Multimedia*, 2010.

[95] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In *Machine learning-international workshop then conference-*, 2001.

[96] C. Wang, S. Yan, L. Zhang, and H.-J. Zhang. Multi-label sparse coding for automatic image annotation. In *CVPR*, 2009.

[97] J. Wang, S. Kumar, and S.-F. Chang. Sequential projection learning for hashing with compact codes. In *ICML*, 2010.

[98] J. Wang and J. Zhu. On statistical analysis and optimization of information retrieval effectiveness metrics. In *SIGIR*, 2010.

[99] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. Annosearch: Image auto-annotation by search. In *CVPR*, 2006.

[100] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, 2005.

[101] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, 2009.

[102] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *JMLR*, 2010.

[103] J. Weston, S. Bengio, and N. Usunier. WSABIE: scaling up to large vocabulary image annotation. In *IJCAI*, 2011.

[104] L. A. Wolsey. *Integer programming*. 1998.

[105] M. Wright. The interior-point revolution in optimization: history, recent developments, and lasting consequences. *Bulletin of the American mathematical society*, 2005.

[106] Y. Wu. *Streaming techniques for statistical modeling*. ProQuest, 2007.

[107] H. Yu, I. Ko, Y. Kim, S. Hwang, and W.-S. Han. Exact indexing for support vector machines. In *SIGMOD*, 2011.

[108] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon. Large-scale multi-label learning with missing labels. In *ICML*, 2014.

[109] Y. Yue and T. Joachims. Predicting diverse subsets using structural svms. In *ICML*, 2008.

[110] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR*, 2003.