Sketches with Unbalanced Bits for Similarity Search

Vladimir Mic(✉), David Novak, and Pavel Zezula

Masaryk University, Brno, Czech Republic
xmic@fi.muni.cz

Abstract. In order to accelerate efficiency of similarity search, compact bit-strings compared by the Hamming distance, so called sketches, have been proposed as a form of dimensionality reduction. To maximize the data compression and, at the same time, minimize the loss of information, sketches typically have the following two properties: (1) each bit divides datasets approximately in halves, i.e. bits are *balanced*, and (2) individual bits have low pairwise *correlations*, preferably zero. It has been shown that sketches with such properties are minimal with respect to the retained information. However, they are very difficult to index due to the dimensionality curse – the range of distances is rather narrow and the distance to the nearest neighbour is high. We suggest to use sketches with *unbalanced* bits and we analyse their properties both analytically and experimentally. We show that such sketches can achieve practically the same quality of similarity search and they are much easier to index thanks to the decrease of distances to the nearest neighbours.

1 Introduction

Treating data objects according to their pairwise similarity closely corresponds to the human perception of reality, thus it represents an important field of data processing. Features of complex objects are typically characterized by *descriptors*, which are often high dimensional vectors. These descriptors can be bulky and evaluation of their pairwise similarity may be computationally demanding [16,21]. Thus techniques to process them efficiently are needed. In this paper we consider one to one mapping between objects and descriptors, thus we do not distinguish these terms and we use just term *object*. One of the state-of-the-art approaches allowing to search big datasets efficiently is based on object transformation to short binary strings – *sketches*. The objective of a *sketching technique* is to construct the binary strings so that they, together with *Hamming distance h*, preserve pairwise similarity relations between objects as much as possible. Thanks to their compact size and computational efficiency of the Hamming distance, sketches have been used by several authors who report promising results for different data types, dimensions, and similarity functions [5,7,15,19].

Many sketching techniques were proposed and majority of them produce sketches $sk(o)$ with *balanced* bits with *low correlations* [12,13,15,19], because these properties are reported to support the quality of similarity approximation:

C. Beecks et al. (Eds.): SISAP 2017, LNCS 10609, pp. 53–63, 2017.
DOI: 10.1007/978-3-319-68474-1_4

- Bit i is balanced (with respect to dataset X) iff it is set to 1 in one half of all sketches $sk(o), o \in X$.
- Bit correlations are investigated in pairwise manner over all pairs of bits of sketches $sk(o), o \in X$.

To the best of our knowledge, there is no prior work discussing disadvantages arising from these properties. In this paper, we analyse their pros and cons and we further focus on sketches with bits balanced to a given ratio b:

- Bit i is balanced to ratio b (with respect to dataset X) iff it is set to 1 in $b \cdot |X|$ sketches $sk(o), o \in X$. Without loss of generality, we assume $0.5 \leq b \leq 1$, since the opposite case is symmetric.

We denote S_b the set of all sketches $sk(o), o \in X$ with bits balanced to b.

We show that the Hamming distance distribution on sketches $S_{0.5}$ (i.e. with balanced bits) with low pairwise bit correlations makes an efficient indexing practically impossible. The main contribution of this paper is analytical and experimental investigation of sketches with unbalanced bits, which shows that they can achieve practically the same quality of the similarity search but they are significantly easier to index.

2 Background

To formalize the concept of similarity, we adopt the model of *metric space* $M = (D, d)$, where D is a domain of objects and $d : D \times D \mapsto \mathbb{R}$ is a *distance function* which determines the dissimilarity of objects [21]. Further we consider a finite dataset $X \subseteq D$. The goal of this section is to provide basic observations about the sketches, which influence their indexability and ability to preserve similarity relationships of objects. First, let us focus on Hamming distance density of sketches S_b with length λ.

Lemma 1 (Mean value of Hamming distance). *Let us have set S_b of sketches $sk(o), o \in X$ with length λ. The mean value of Hamming distance on S_b is $2\lambda \cdot b \cdot (1 - b)$ regardless of pairwise bit correlations.*

Proof. Let us consider one bit i of the sketches. The Hamming distance h of sketches $sk(o_1)$, $sk(o_2)$ on bit i is 1 iff $sk(o_1)$, $sk(o_2)$ have different values in bit i. Considering all $|X|$ sketches, it happens in $2b|X| \cdot (1 - b)|X|$ cases. Thus sum of all Hamming distances over λ bits is $2\lambda b|X| \cdot (1 - b)|X|$, and the mean Hamming distance is $2\lambda b(1 - b)$ as we summed $|X|^2$ distances in the previous step.

The mean of Hamming distance is maximized for $b = 0.5$, i.e. for the balanced bits.

Next, we focus on the bit correlation, which we express with the Pearson correlation coefficient. According to Theorem 2 in [12], the variance of Hamming distance on sketches $S_{0.5}$ decreases with decreasing absolute value of the average pairwise bit correlation; it is minimized for uncorrelated bits.

In case of uncorrelated bits, the Hamming distance density of sketches S_b has binomial distribution, thus for variance σ^2 of Hamming distances holds: $\sigma^2 = \lambda b(1-b)$. In other words, (1) sketches with balanced bits have maximum mean distance and, (2) for these sketches, minimization of the pairwise bit correlations means minimization of the variance of the Hamming distance, which is maximization of all distances lower than the mean distance. Clearly, maximizing values of the smallest inter-object distances violates the key objective of the data transformation for the similarity indexing: distances $h(sk(o_1), sk(o_2)), o_1, o_2 \in X$ for very similar objects o_1, o_2 are desired to be small [5]. Moreover these consequences imply a problem known as the dimensionality curse [18].

A formalised view is provided by the *intrinsic dimensionality* of sketches. Intrinsic dimensionality (*iDim*) expresses "*the minimum number of parameters needed to account for the observed properties of the data*" [6]. We use the formula proposed by Chavez and Navarro [2] for the estimation of *iDim*:

$$iDim \approx \frac{\mu^2}{2 \cdot \sigma^2}, \tag{1}$$

where μ is the mean of distance density, and σ^2 is its variance. In compliance with the previous paragraph, it has been proven that:

- for uncorrelated bits, *iDim* is maximized iff they are balanced [18],
- for balanced bits, *iDim* is maximized iff they are uncorrelated [12].

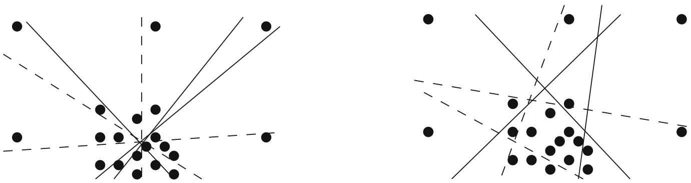
In the field of similarity search, *iDim* expresses "the difficulty" of data indexing [17]. Thus techniques which produce sketches $S_{0.5}$ with bit correlations close to zero produce hard-to-index sketches. Moreover indexing techniques typically assume at least a few objects in small distances from the query object [14,16].

2.1 Observations on GHP Sketches

We illustrate our findings on a sketching technique based on the *generalized hyperplane partitioning* (GHP) [21]. Bit i of all sketches $sk(o), o \in X$ is determined using a pair of pivoting objects p_{i0}, p_{i1}, which splits objects $o \in X$ by comparing distances $d(o, p_{i0}), d(o, p_{i1})$; value of bit $sk_i(o)$ expresses which pivot is closer to o. This technique is described in detail e.g. in [12].

Let us consider query object $q \in D$ and its most similar object $o_{q1} \in X$, $o_{q1} \neq q$. As we have explained, the Hamming distance $h(sk(q), sk(o_{q1}))$ is high on average on sketches $S_{0.5}$ with low pairwise bit correlations. It means that many hyperplanes separate q and o_{q1}. On the contrary, in case of sketches with unbalanced bits, e.g. $S_{0.8}$, the distance $h(sk(q), sk(o_{q1}))$ should be lower.

For motivation, consider the situation in 2D Euclidean space shown in Fig. 1. In case of hyperplanes dividing dataset into halves (Fig. 1a), the Hamming distances between originally close objects are high which suggests that many hyperplanes split dense subspaces of dataset X. On the other hand, in case of sketches with unbalanced bits (Fig. 1b), the Hamming distances are smaller not only for



(a) $b = 0.5$

(b) $b = 0.8$

Fig. 1. Hyperplanes producing sketches with bits balanced to different balance b

originally close objects, but also for more distant ones; as shown later, this drawback can be compensated by using longer sketches. Please, note that this is only an artificial example in 2D, but these properties are implied by values of mean and variance of the Hamming distance and thus hold even for real-life, high dimensional data.

Figure 1b suggests, that unbalanced bits may lead to many objects with all bits set to 1 (the "center" of the figure). However, our practical experience with sketches in high dimensional space show that there is just a few such objects. In particular, we conducted experiments with $b \in \{0.85, 0.9\}$ and $\lambda = 205$ (see Sect. 4 for details on the dataset). We realized that there was no sketch with all bits set to 1 in case of $b = 0.85$, and only eight out of one million in case of $b = 0.9$.

2.2 Related Work

Charikar has introduced in his pioneering work [1] the idea of using random hyperplanes to summarize objects in a multi-dimensional vector spaces in such a way that the resulting bit strings respect the cosine distance. Lv et al. [11] have proposed a sketching technique for spaces of vectors compared by (weighted) L_1 distance function. Their method is based on *thresholding*; a threshold is determined for each dimension of the original space. Individual bits of the sketches are set according to values in corresponding dimensions and compressed. Pagh et al. propose *odd sketches* [14] – short binary strings created as a transformation of original vector space, based on *min-hashing* [9,10]. Odd sketches are compared by the Hamming distance, and they suffer a lot from the curse of dimensionality. Daugman [3] uses bit strings to describe human irises to identify people. His method is based on encoding shades of grey colour shown in images of irises in UV light, and it constitutes the most widely used approach of human irises recognition.

3 Analysis of Bits Balanced to *b*

The objective of this section is to quantify all trends mentioned above. More specifically, we analytically derive the influence of the ratio b on bit correlations, Hamming distance distribution, and *iDim* of sketches.

Consider sketches S_b and their two arbitrarily selected bits i, j, $0 \leq i < \lambda$, $0 \leq j < \lambda$. Set S_b can be split into four parts according to combination of values in bits i, j. Let us denote $\#_{11}, \#_{10}, \#_{01}, \#_{00}$ the relative numbers of the sketches in these four parts. It holds that $\#_{10} = \#_{01}$ regardless of correlation of bits i and j. Denoting $sk_i(o)$ the ith bit of sketch $sk(o)$, the Pearson correlation coefficient of bits i, j can be simplified:

$$Corr(i,j) = \frac{\sum\limits_{o \in X} (sk_i(o) - b)(sk_j(o) - b)}{\sqrt{\sum\limits_{o \in X} (sk_i(o) - b)^2 \sum\limits_{o \in X} (sk_j(o) - b)^2}} = \frac{\#_{11} - b^2}{b(1-b)}. \tag{2}$$

Let us point out one difference between balanced and unbalanced bits: if we switch all values in arbitrary bit i in case of balanced bits, only the sign of correlations $Corr(i, j)$ (with all other bits j) changes. Thus only the absolute values of pairwise bit correlations matter [12]. On the other hand, in case of unbalanced bits ($b \neq 0.5$), the sign of correlation matters, as opposite correlations express different space partitioning. For example for $b = 0.8$, correlation -0.25 means object distribution $\#_{11} = 60\,\%$, $\#_{10} = \#_{01} = 20\,\%$, $\#_{00} = 0\,\%$ while correlation $+0.25$ means distribution $\#_{11} = 68\,\%$, $\#_{10} = \#_{01} = 12\,\%$, $\#_{00} = 8\,\%$. Therefore, we keep the same bit orientation for all bits (specifically, 1 in $b \cdot |X|$ sketches).

It has been shown [12], that high intrinsic dimensionality *iDim* increases potential of sketches with balanced bits to well approximate similarity relationships of objects. So, let us analyse the *iDim* of sketches with bits balanced to b. We denote H_i the list of all $|X|^2$ Hamming distances measured just on bit i of sketches $sk(o), o \in X$. Then $Corr(H_i, H_j)$ is the Pearson correlation of lists H_i, H_j, and $Corr_{Avg}$ is the average pairwise correlation over all lists $H_i, H_j, 0 \leq i < j < \lambda$. We have derived in [12] the variance σ^2 of Hamming distance on sketches $S_{0.5}$. Using Lemma 7 from that paper and analogous approach, it is possible to derive σ^2 for sketches S_b:

$$\sigma^2 = 2b(1-b) \cdot [1 - 2b(1-b)] \cdot [\lambda + (\lambda^2 - \lambda) \cdot Corr_{Avg}]. \tag{3}$$

Using Lemma 1 and Eq. 3, the *iDim* of sketches with bits balanced to b is:

$$iDim \approx \frac{\mu^2}{2\sigma^2} = \frac{b \cdot (1-b) \cdot \lambda}{(2b^2 - 2b + 1) \cdot [1 + (\lambda - 1)\, Corr_{Avg}]}. \tag{4}$$

Therefore *iDim* of sketches increases with decreasing $Corr_{Avg}$. In the following, we lower bound average correlation $Corr_{Avg}$, which implies the upper bound for *iDim* of sketches with bits balanced to b.

Minimum average correlation $Corr_{Avg}$ occurs iff all pairwise correlations of lists H_i, H_j are minimal. Thus, we focus on $Corr(H_i, H_j)$:

$$Corr(H_i, H_j) = \frac{2(\#_{11} \cdot \#_{00} + \#_{10}^2) - [2b(1-b)]^2}{2b(1-b) - [2b(1-b)]^2}. \tag{5}$$

Using this equation, it is possible to express values $\#_{11}, \#_{10}$ and $\#_{00}$ implying minimum value of $Corr(H_i, H_j)$. Please, notice that all fractions $\#_{11}, \#_{10}$ and $\#_{00}$ must be non-negative.

Theorem 1. *Maximum iDim of sketches S_b with bits balanced to b occurs iff for all pairs of bits $0 \leq i < j < \lambda$ holds: $\#_{00} = \max(0, 3/4 - b)$, $\#_{10} = \min(1/4, 1-b)$ and $\#_{11} = \max(2b-1, b-1/4)$.*

Proof. Theorem holds as a consequence of Eqs. 2, 4 and 5.

Values $\#_{00}$,$\#_{10}$ and $\#_{11}$ implying maximum *iDim* of sketches, imply negative pairwise correlations $Corr(i, j)$ for $b > 0.5$, which bring a problem: it is not possible to create meaningful sketches for similarity search with significantly negative pairwise bit correlations. Considering a given ratio $b > 0.5$ and negative bit correlations, each zero in an arbitrary bit i of any sketch $sk_i(o_1)$ pushes other values $sk_i(o_2), o_2 \in X \wedge o_2 \neq o_1$ to be 1. However the number of ones is given by ratio b.

In case of $b \geq 0.75$, maximum *iDim* occurs iff $\forall 0 \leq i < j \leq \lambda : \#_{00} = 0$. In this case, each sketch contains exactly one or none bit set to 0, and therefore at most $\lambda + 1$ different sketches of length λ exist (including one with all bits set to 1). In the other words, an effort to minimize ratio $\#_{00}$ leads to extremely long sketches.

In practice when a realistic sketch length λ is preserved, higher *iDim* of sketches S_b may be achieved with an effort to produce uncorrelated bits rather than negatively correlated. The reason is, that few significant negative correlations usually cause higher increase of other correlations which leads to an increase of average pairwise correlation above zero. We illustrate these statements in an experiment in Sect. 4.

As a result of provided analysis and experiments, we propose to search for uncorrelated unbalanced bits, i.e. for sketches with binomial distance distribution, but with lower mean value than in case of balanced bits, which is favourable for indexing of sketches.

4 Evaluation

At first, we run an experiment to confirm the suitability of producing sketches with uncorrelated rather than negatively correlated bits. Then we focus on the quality of similarity search with unbalanced sketches and their indexability. Let us briefly describe the testing data and sketching technique:

Testing Data

The experiments are conducted on a real-life data collection consisting of visual descriptors extracted from images. More specifically, we use *DeCAF* descriptors [4] – 4096-dimensional vectors taken as an output from the last hidden layer of a deep convolutional neural network [8]. These descriptors were extracted from a 1M subset of the *Profiset collection*[1]. The DeCAF descriptors are compared by the Euclidean distance to form a metric space.

Sketching Technique

In order to create sketches, we randomly select a set of 512 pivots and we investigate all $\binom{512}{2}$ pivot pairs. We use a random subset of 100,000 data objects and analyse the balance b of generalized hyperplane partitioning (GHP) defined by each pair of pivots (see Sect. 2.1 for examples of GHP). From pivot pairs implying a proper balance b (which is about 8,000–15,000 pairs) we further select those, producing sketches with low correlated bits using our heuristic. Description of this heuristic is available online[2].

4.1 Searching for Negatively Correlated Bits

Table 1 contains evaluated properties of sketches created by the sketching technique, which tried to (1) find sketches with uncorrelated bits, and (2) find as negatively correlated bits as possible. Ratio b was 0.8 in these experiments, and results for four sketch lengths λ are presented. The average pairwise bit correlation is lower in case of searching for uncorrelated bits, rather then negatively correlated, in three cases. The numbers of negative and positive pairwise bit correlations confirms these results as well. There is an exception in Table 1, the sketch length $\lambda = 205$ for which average bit correlation is lower when searched for negatively correlated bits. However, observed difference is tiny in this case.

Table 1. Sketching technique: searching for uncorrelated and negatively correlated unbalanced bits, $b = 0.8$

Searching for uncorrelated				Searching for negative correlations		
λ	Average corr	# positive	# negative	Average corr	# positive	# negative
64	+0.0019	1,000	1,016	+0.0024	1,032	984
128	+0.0046	4,066	4,062	+0.0053	4,106	4,022
205	+0.0064	10,494	10,416	+0.0063	10,469	10,441
256	+0.0072	16,406	16,234	+0.0077	16,436	16,204

The reasons of observed tendencies are discussed in theoretical Sect. 3.

[1] http://disa.fi.muni.cz/profiset/.

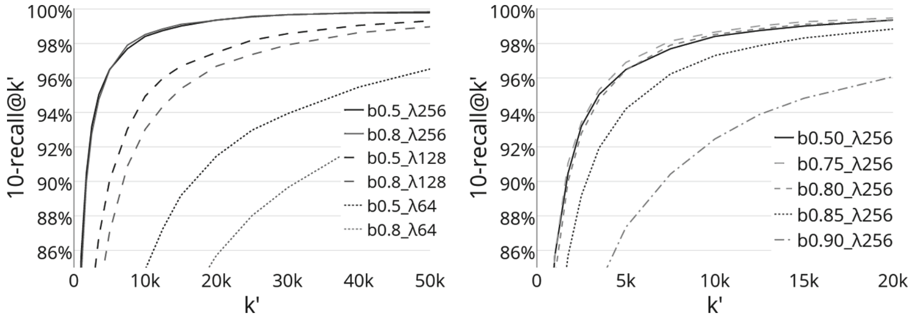
[2] http://www.fi.muni.cz/~xmic/sketches/AlgSelectLowCorBits.pdf.

4.2 Quality of Sketches

The most important requirement for sketches with unbalanced bits is that they have to provide acceptable quality of the similarity search in comparison to sketches with balanced bits. In the following experiments, we use k-recall@k' of approximate kNN search using sketches. More specifically, for each query object q, we compare the set of k most similar objects from X found by the sequential scan of X (denoted as $Prec(q)$) with k objects found by the *filter and refine* approach based on sketches: First, in the filtering phase, we select k' objects $o \in X$, $k' \geq k$ with smallest Hamming distances $h(sk(q), sk(o))$. Then these k' objects o are refined by evaluating distances $d(q, o)$ in order to identify approximate kNN answer denoted as *Ans*(*q, k'*). The ability of sketches to approximate similarity relationships of objects $o \in X$ is expressed by measure k-recall@k':

$$k\text{-recall@}k' = \frac{Prec(q) \cap Ans(q, k')}{k}. \tag{6}$$

In the following, we present results only for $k = 10$, because trends observed in these experiments are the same even for other values of k. Size of dataset X in the following experiments is $|X| = 1{,}000{,}000$. All results are averages over 1,000 randomly selected queries q.



(a) $b \in \{0.5, 0.8\}, \lambda \in \{64, 128, 256\}$ (b) $b \in \{0.5, 0.75, 0.8, 0.85, 0.9\}, \lambda = 256$

Fig. 2. Quality of approx. similarity search with sketches balanced to different b

We demonstrate by Fig. 2a that the difference in 10-recall@k' using sketches $S_{0.5}$ and $S_{0.8}$ is relatively high in case of short sketches, however with increasing length λ it is becoming negligible (in our case, this happens approximately for $\lambda \geq 200$). Figure 2b depicts 10-recall@k' for sketch length $\lambda = 256$ and different ratio b. Using $b \in \{0.5,\ 0.75,\ 0.8\}$, the results are practically the same; decrease is noticeable in case of $b = 0.85$: for example about 2.3 percentage points for k' = 5,000 (i.e. 0.5 % of $|X|$) and it is significant for $b = 0.9$: e.q. 9.2 percentage points for k' = 5,000.

Table 2. Sketches with $\lambda = 256$: 10-recall@k', *iDim* and avg. Hamming distances to k'th closest sketch

b	10-recall@k'		*iDim*	$h(sk(q), sk(o_{qk'}))$		
	k'=2,500	k'=10,000		k'=1	k'=100	k'=10,000
0.50	93.20 %	98.41 %	29.4	43.1	58.9	83,6
0.75	93.43 %	98.66 %	24.1	32.0	44.3	63.2
0.80	92.79 %	98.52 %	19.6	27.4	38.0	54.0
0.85	89.19 %	97.30 %	13.5	21.1	29.7	42.2
0.90	80.31 %	92.46 %	9.0	13.8	19.9	28.3

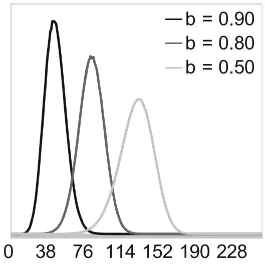


Fig. 3. Ham. distance densities for $\lambda = 256$

4.3 Indexability of Sketches

The indexability of sketches is illustrated by their *iDim* and by the average Hamming distances $h(sk(q), sk(o_{qk'}))$ between $sk(q)$ and its k'th nearest sketch for $k' \in \{1;\ 100;\ 10{,}000\}$. We show results for $b \in \{0.5, 0.75, 0.8, 0.85, 0.9\}$ in Table 2. These results make possible to utilize techniques for bit-strings indexing and other processing [16,20]

As expected, the *iDim* of sketches decreases as ratio b grows (for $b \geq 0.5$). For instance the *iDim* of sketches $S_{0.5}$ and $S_{0.8}$ differs about one third for $\lambda = 256$. In order to remind results from Sect. 4.2, we show 10-recall@k' for two selected k': the difference of 10-recall@k' for $b \in [0.5,\ 0.8]$ is negligible. It confirms, that properly unbalanced sketches can be used as a full-fledged but easily indexable alternative to sketches with balanced bits. Better indexability is confirmed by the decrease of distances to the k' nearest sketches (shown in last three columns of Table 2), and by distribution of Hamming distance densities presented in Fig. 3. All these measurements confirm the analytic results from Sects. 2 and 3.

5 Conclusions

We have investigated sketches – bit strings created by such transformation of data objects, which should preserve the similarity relationships between the objects. Sketching techniques proposed so far usually aim at producing bit strings with balanced and low correlated bits. Sketches with these properties have been reported to provide the best trade-off between their length and ability to approximate similarity relationships between objects. In this paper, we studied one drawback of such sketches: these properties lead to maximization of the *intrinsic dimensionality* of the set of sketches making them hard-to-index (because of the *dimensionality curse*). We thus focus on sketches with bits balanced to some given ratio b and we derive various theoretical properties of such sketches. Further, we show on a real life dataset that the proposed approach can achieve practically the same quality of the similarity search, but with sketches having *iDim* about one third lower than sketches with balanced bits.

Acknowledgements. We thank to Matthew Skala for his advices about negative pairwise bit correlations. This research was supported by the Czech Science Foundation project number GBP103/12/G084.

References

1. Charikar, M.S.: Similarity estimation techniques from rounding algorithms. In: Proceedings of the 34th Annual ACM Symposium on Theory of Computing. ACM, New York (2002)
2. Chávez, E., Navarro, G., Baeza-Yates, R., Marroquín, J.L.: Searching in metric spaces. ACM Comput. Surv. **33**(3) (2001)
3. Daugman, J.: The importance of being random: statistical principles of iris recognition. Pattern Recognit. **36**(2) (2003)
4. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: a deep convolutional activation feature for generic visual recognition. In: ICML 2014, vol. 32, pp. 647–655 (2014)
5. Dong, W., Charikar, M., Li, K.: Asymmetric distance estimation with sketches for similarity search in high-dimensional spaces. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (2008)
6. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, San Diego (2013)
7. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 304–317. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88682-2_24
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
9. Leskovec, J., Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press, Cambridge (2014)
10. Li, P., König, A.C.: Theory and applications of b-bit minwise hashing. Commun. ACM **54**(8), 101–109 (2011)
11. Lv, Q., Charikar, M., Li, K.: Image similarity search with compact data structures. In: Proceedings of the 13th ACM International Conference on Information and Knowledge Management, pp. 208–217. ACM (2004)
12. Mic, V., Novak, D., Zezula, P.: Designing sketches for similarity filtering. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 655–662, December 2016
13. Mic, V., Novak, D., Zezula, P.: Speeding up similarity search by sketches. In: Amsaleg, L., Houle, M.E., Schubert, E. (eds.) SISAP 2016. LNCS, vol. 9939, pp. 250–258. Springer, Cham (2016). doi:10.1007/978-3-319-46759-7_19
14. Mitzenmacher, M., Pagh, R., Pham, N.: Efficient estimation for high similarities using odd sketches. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 109–118. ACM (2014)
15. Muller-Molina, A.J., Shinohara, T.: Efficient similarity search by reducing i/o with compressed sketches. In: Proceedings of the 2nd International Workshop on Similarity Search and Applications, pp. 30–38 (2009)

16. Pagh, R.: Locality-sensitive hashing without false negatives. In: Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1–9. Society for Industrial and Applied Mathematics (2016)
17. Skala, M.: Measuring the difficulty of distance-based indexing. In: Consens, M., Navarro, G. (eds.) SPIRE 2005. LNCS, vol. 3772, pp. 103–114. Springer, Heidelberg (2005). doi:10.1007/11575832_12
18. Skala, M.A.: Aspects of Metric Spaces in Computation. Ph.D. thesis, University of Waterloo (2008)
19. Wang, Z., Dong, W., Josephson, W., Lv, Q., Charikar, M., Li, K.: Sizing sketches: a rank-based analysis for similarity search. SIGMETRICS Perform. Eval. Rev. **35**(1), 157–168 (2007)
20. Zezula, P., Rabitti, F., Tiberio, P.: Dynamic partitioning of signature files. ACM Trans. Inf. Syst. **9**(4), 336–367 (1991)
21. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search: The Metric Space Approach, vol. 32. Springer, Boston (2006)