Selecting Sketches for Similarity Search

Vladimir Mic[1], David Novak[1], Lucia Vadicamo[2] and Pavel Zezula[1]

[1] Masaryk University
Brno, Czech Republic

[2] CNR-ISTI
Pisa, Italy

4th September 2018

Similarity Search

- Field: searching for similar objects
- *Queries by example*
  - The goal is to efficiently find the most similar objects to a given query object

- Wide range of applications
  - information retrieval, recommender systems, searching in biometrics, event detection, ...

Formalization

- We consider similarity modelled by the metric space (D, d)
  - D – domain of objects – original objects or their descriptive features
  - $d : D \times D \mapsto \mathbb{R}^+$ – distance function
    - the bigger the value $d(o_1, o_2)$, the less similar objects o_1, o_2
- Dataset $X \subseteq D$
- Having a query object $q \in D$, the goal is to efficiently find the most similar objects $o \in X$ to q

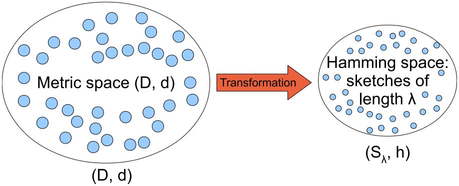
Challenges

- Challenges:
  - Dataset X usually contains a lot of objects
  - Objects $o \in X$ are often big
  - Similarity function d may be complex and expensive

  - We have limited computational power
  - Queries have to be evaluated fast

Bit String Sketches

- Successful family of techniques to mitigate these problems: transformation of the metric space (D, d) to the Hamming space
  - sketch $sk(o)$ of object $o \in X$ is bit string of length λ
  - $sk : D \mapsto \{0, 1\}^{\lambda}$: *sketching* (transformation) *technique*
- Sketches compared by the Hamming distance approximate similarity relationships between objects $o \in X$



- Sketches are small ($\lambda \approx 64 - 256$ bits)
- Evaluation of the Hamming distance is efficient

Issues and Challenges

- Current state
  - many sketching techniques *sk* exist
  - each sketching technique is suitable for just some datasets
  - sketch length λ and all parameters of *sk* must be set, which requires an expert knowledge or complex testing

Our Objectives

- we provide a tool to efficiently estimate a quality of a sketching technique *sk* considering a given dataset

Testing Sketching Techniques

- Established way of testing is expensive:
  - select a sample set of X of a representative size
  - select a set of query objects $Q \subseteq D$
- and compare precise query results
  - k most similar objects $o \in X$ to each query $q \in Q$
- with approximate (and more efficient) query evaluation based on sketch filtering
- This comparison is made for each investigated sketching technique to select the best one

Testing Sketching Techniques

- Established way of testing is expensive:
  - select a sample set of X of a representative size
  - select a set of query objects $Q \subseteq D$
- and compare precise query results
  - k most similar objects $o \in X$ to each query $q \in Q$
- with approximate (and more efficient) query evaluation based on sketch filtering
  - identify $c \geq k$ most similar sketches $sk(o)$ to $sk(q)$
  - access objects $o \in X$ that correspond to the most similar sketches (*candidate objects*) and evaluate distances $d(q, o)$
  - answer: k most similar candidate objects
- This comparison is made for each investigated sketching technique to select the best one

Pros and Cons of Established Testing

- This established testing:
  - is affected by a selection of query objects Q
  - dataset X and query set Q must be of sufficient (big) size
    - we usually use $|X| \geq 1{,}000{,}000$ objects
  - all sketches for each $o \in X$ and $q \in Q$ must be created
  - evaluation of precise answers for each query object $q \in Q$ must be performed (it is expensive)
  - quality of approximate evaluations is strongly influenced by the number of selected candidate objects c
  - selecting c with no prior knowledge of the sketching technique is difficult

  - therefore: very expensive procedure with limited detachment

  + comparison of precise and approximate answer is intuitive and easy to understand

Our Contribution

- We propose two efficient methods to estimate quality of sketches $sk(o)$, $o \in X$
  - i.e., their ability to approximate similarity relationships of objects $o \in X$
  - Both use just a very small sample set of data
- Both are based on probabilistic analysis

Pros and Cons of Our Methods

- Our methods
  - $+$ do not use any query objects Q (so are not affected by their selection)
  - $+$ small sample set of $\approx$ 2,000 – 5,000 objects $o \in X$ is sufficient for our estimations
  - $-$ all sketches $sk(o)$ for the sample set must be created
  - $+$ no need to expensively evaluate any *precise query answers*
  - $+$ no candidate set is used, so no expert knowledge or testing to set its size is required
  - $+$ therefore: efficient methods, easy to use
    - Examination of a set of sketches made by a given sketching technique sk requires less than 1 minute
  - $-$ quality of sketching technique is expressed by an abstract real number with *no intuitive meaning*

Our Approach

- Let us have a sketching technique sk producing sketches of length λ and distance $x = d(o_1, o_2)$
- we model[1] probability $p(x, b)$ that the Hamming distance of sketches $sk(o_1), sk(o_2)$ is b for $0 \leq b \leq \lambda$, i.e. $h(sk(o_1), sk(o_2)) = b$

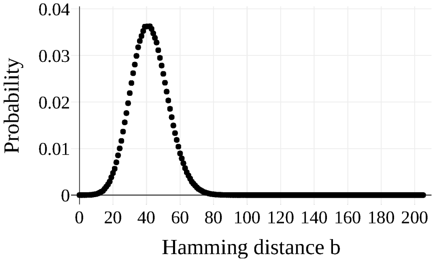


Figure: Example of probability function $p(x, b)$ for a given value x

[1] details later

Projection of Two Distances x_1, x_2

- Consider two distances $x_1 < x_2$ and functions $p(x_1, b)$, $p(x_2, b)$

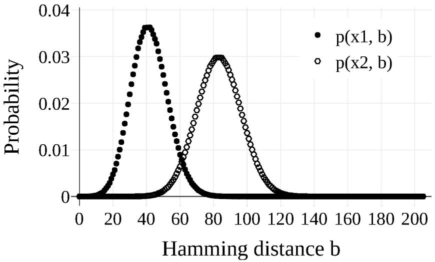


Figure: Functions $p(x_1, b)$, $p(x_2, b)$ for given values x_1, x_2

- Ideal case: sketching technique preserve ordering of distances
  - i.e. $x_1 < x_2 \implies h(sk(o_1), sk(o_2)) < h(sk(o_3), sk(o_4))$
- We evaluate separation of probability functions $p(x_1, b)$, $p(x_2, b)$

- m_1, m_2 ... means of $p(x_1, b)$, $p(x_2, b)$
- s_1^2, s_2^2 ... variances of $p(x_1, b)$, $p(x_2, b)$

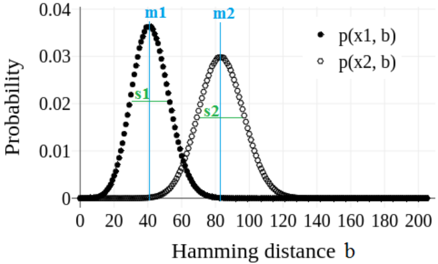


Figure: Functions $p(x_1, b)$, $p(x_2, b)$ for given values x_1, x_2

- Separation of functions[2]

$$sep_{sk}(x_1, x_2) = \frac{m_2 - m_1}{\sqrt{\frac{s_1^2 + x_2^2}{2}}}$$

[2] adopted formula

Quality of Sketching Technique

- Quality of a sketching technique sk:
  We evaluate $sep_{sk}(x_1, x_2)$ over whole range $[0, \Gamma]$ of distances x_1, x_2:

$$quality(sk) = \int_0^{\Gamma} \int_{x_1}^{\Gamma} sep_{sk}(x_1, x_2)\, \partial x_2\, \partial x_1$$

Interpretation

Value *quality(sk)* describes, how much a sketching technique sk distinguishes distances between objects $o \in X$, i.e. quality of sk

- Possible modifications:
  (1) normalization by Γ^2
  (2) similarity search: focus on separation of *small distances* (that are smaller than some t) from others

$$quality_{norm}(sk,\ t) = \frac{\int_0^{t} \int_{x_1}^{\Gamma} sep_{sk}(x_1, x_2)\, \partial x_2\, \partial x_1}{\Gamma^2}$$

Approaches to model function $p(x, b)$

- Details: two approaches to model probability function $p(x, b)$
  - Approach A (*analytique*)
  - Approach PM (*partially measured*)

- Both approaches use
  - set of distances $d(o_1, o_2)$ and
  - corresponding Hamming distances on sketches $h(sk(o_1), sk(o_2))$

  to estimate means m and variances s^2 of $p(x, b)$, and therefore $sep_{sk}(x_1, x_2)$:

$$sep_{sk}(x_1, x_2) = \frac{m_2 - m_1}{\sqrt{\frac{s_1^2 + x_2^2}{2}}}$$

Approach A

- Approach A models (complete) function $p(x, b)$
  - precomputed distances are investigated to get an average probability $p_i(x, 1)$ that one bit of sketches $sk(o_1)$ and $sk(o_2)$ is different
  - complete $p(x, b)$ is modelled by a composition of λ instances of $p_i(x, 1)$
  - Approach A reveals statistical properties of sketches that improve their quality
- Approach PM:
  - means and variances m and s^2 of $p(x, b)$ are directly evaluated using precomputed distances d and h
  - Approach PM does not reveal statistical properties of sketches that improve their quality

Experiments – Description

We experimentally verify our estimators by their comparison with the established testing procedure

- 4 different sketching techniques *sk*
  - based on *generalyzed hyperplane partitioning* (*GHP50*, *GHP80*), *ball partitioning* (*BP50*), and *thresholding* (*THRR50*)
  - their detailed description is in the paper
- For each technique 4 different lengths λ are examined (if possible)
  - 64, 128, 192, 256 bits
- Two datasets of size $|X| = 1,000,000$ vectors, each
  - real-valued vectors of length 4,096 (*DeCAF from neural network*)
  - real-valued vectors of length 128 (*SIFT: local visual image descriptors*)

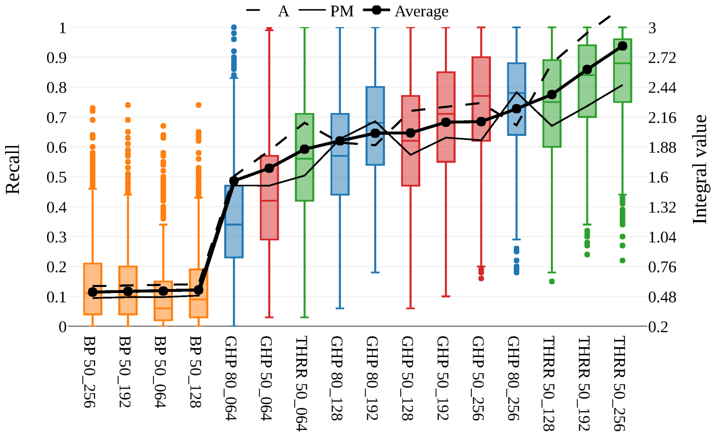
- Established testing procedure:
  - *the recall value*: size of intersection of the precise and approximate query answers
  - 1,000 queries q, search for 100 nearest neighbour
  - candidate set size: 2,000 objects (i.e. 0.2 %)

  - Costs:
    - Precise answers: up to 2 billion $d(q, o)$ evaluations (brute force)
    - among other things, 6.5 billion $d(o_1, o_2)$ evaluations to create 30 different sets of sketches

Experiments – Our Estimations

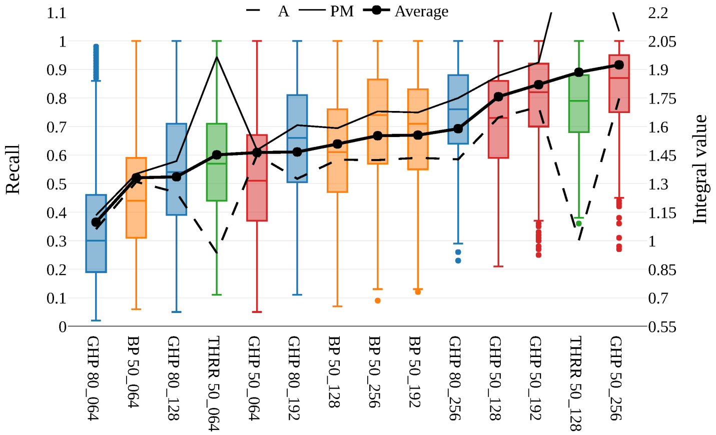
- Our estimators use 5,000 randomly selected objects $o \in X$ and their sketches $sk(o)$ made by each investigated sketching technique
- We evaluate 2,000,000 distances $d(o_1, o_2)$ and corresponding $h(sk(o_1), sk(o_2))$ to get our estimations
- Estimation takes 30 – 50 seconds per set of sketches

Results – DeCAF dataset



- *x*-axis: sets of sketches, 3 last digits: sketch length λ, colours of box plots: principaly different sketching techniques
- primary *y*-axis: the recall examined by expensive *established testing* (box plots)

Results – SIFT dataset



- x-axis: sets of sketches, 3 last digits: sketch length λ, colours of box plots: principaly different sketching techniques
- primary y-axis: the recall examined by expensive *established testing* (box plots)

Results – Correlations

- Average of both estimations – high quality results (possible since both estimations use the same scale)

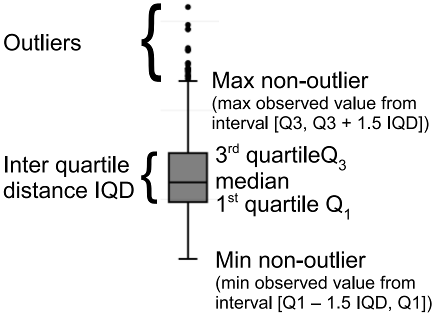
Table: Correlations of quality estimations and measured medians of the recall

	Approach A	Approach PM	Average of estimations
DeCAF	+0.96	+0.97	+0.98
SIFT	+0.55	+0.74	+0.93

- Conclusions:
  - We proposed analytical tools to estimate quality of binary sketches
  - They use very small sample of data
  - They are very efficient

Results

- The recall value (i.e. quality of sketch based filtering examined by the established approach) is expressed by box plots to show distribution of values among particular query objects $q \in Q$



Contribution of Sketches

Question in reviews: how about scalability of sketches:

- If sketches are not indexed[3], just their *quality* matters
- Indexing of sketches is hard, in general, due to big Hamming distance to nearest neighbours
- *Indexability* of sketches made a given sketching technique *sk* strongly depends on a dataset
- We cannot make conclusions about the examined techniques based on testing on 2 datasets ...

[3] i.e. our case: sequential evaluation of (all) Hamming distances is considered