Efficient Similarity Search in Scientific Databases with Feature Signatures

Merih Seran Uysal  Christian Beecks  Jochen Schmücking  Thomas Seidl

RWTH Aachen University, Germany

27th International Conference on Scientific and Statistical Database Management
July 1, 2015
San Diego, California
Similarity Search

How to determine the similarity between two data objects in scientific databases?

Requirements

- Feature representation
- Similarity measure
- Efficient query processing
Similarity Search

How to determine the similarity between two data objects in scientific databases?

Requirements

- Feature representation $\Rightarrow$ use feature signatures
- Similarity measure $\Rightarrow$ use Earth Mover’s Distance
- Efficient query processing $\Rightarrow$ new lower-bounding techniques
Overview

1. Preliminaries
2. Reduced Signatures
3. Filter Approximations
4. Experimental Evaluation
5. Conclusion and Outlook
Feature Representation

Represent the data object by features in a feature space $F$

Aggregate features to obtain a compact feature representation

Finite number of features with non-zero weights ($representatives$)

Formally; $X : F \rightarrow \mathbb{R}$ subject to $|R_X| < \infty$ with $R_X = \{ f \in F | X(f) \neq 0 \} \subseteq F$
Intuition: Earth Mover’s Distance

Earth Mover’s Distance (EMD) [1]
- transforms each feature signature to another one
- denotes a transportation problem (linear optimization problem)
- chooses the minimum-cost flow among all flows
- exhibits high computational time complexity

Earth Mover’s Distance

Given feature signatures $X, Y \in \mathbb{S}^+$ over a feature space $(\mathbb{F}, \delta)$ with a distance function $\delta : \mathbb{F} \times \mathbb{F} \to \mathbb{R}$, $\text{EMD} : \mathbb{S}^+ \times \mathbb{S}^+ \to \mathbb{R}$ between $X$ and $Y$ is defined as a minimum-cost flow of all possible flows $F = \{f : \mathbb{F} \times \mathbb{F} \to \mathbb{R}\} = \mathbb{R}^{\mathbb{F} \times \mathbb{F}}$ by:

$$\text{EMD}(X, Y) = \min_{f \in F} \left\{ \frac{1}{m} \sum_{x \in \mathbb{F}} \sum_{y \in \mathbb{F}} \delta(x, y) \cdot f(x, y) \right\}$$

subject to constraints:

- Non-negativity: $\forall x, y \in \mathbb{F} : f(x, y) \geq 0$
- Source: $\forall x \in \mathbb{F} : \sum_{y \in \mathbb{F}} f(x, y) \leq X(x)$
- Target: $\forall y \in \mathbb{F} : \sum_{x \in \mathbb{F}} f(x, y) \leq Y(y)$
- Total flow: $m = \sum_{x \in \mathbb{F}} \sum_{y \in \mathbb{F}} f(x, y) = \min \{ \sum_{x \in \mathbb{F}} X(x), \sum_{y \in \mathbb{F}} Y(y) \}$
High computational time complexity of the EMD ⇒ Bottleneck!
How to perform query processing with the EMD on signatures **efficiently**?
Efficient Query Processing

High computational time complexity of the EMD ⇒ Bottleneck!
How to perform query processing with the EMD on signatures efficiently?

Approach:
- Utilize reduced signatures via some heuristics
- Lower-bounding filter distance functions on reduced signatures
- Filter-and-refine architecture

Filter:
- Completeness (no false dismissal) ⇒ lower bound property
- Selectivity ⇒ a small candidate set
- Efficiency
Efficient Query Processing

High computational time complexity of the EMD ⇒ Bottleneck!
How to perform query processing with the EMD on signatures efficiently?

Approach:
- Utilize reduced signatures via some heuristics
- Lower-bounding filter distance functions on reduced signatures
- Filter-and-refine architecture

Filter:
- Completeness (no false dismissal) ⇒ lower bound property
- Selectivity ⇒ a small candidate set
- Efficiency
Reduced Signature

Definition (Reduced Signature)

Let $X, X' \in S_{\geq 0}$ be two feature signatures. $X'$ is a reduced feature signature with respect to $X$ if it holds: $\forall x \in \mathbb{F} : X(x) \geq X'(x)$.

- Dimensionality reduction is a special case of signature reduction

(a) EMD flow between two signatures $\Rightarrow$ EMD($X, Y$)$=3.2$
(b) Removing a representative and EMD flow $\Rightarrow$ EMD($X', Y$)$=3.71$

Discarding representatives in a signature leads to a higher EMD value $\Rightarrow$ Completeness not preserved!
Definition (\(\lambda\)-Independent Minimization for Signatures)

Let \((F, \delta)\) be a feature space with a distance function \(\delta\), \(X, Y \in S_{\geq 0}\) be feature signatures, and \(\lambda \in \mathbb{R}^+\). \(\lambda\)-\textsc{IM-Sig} : \(S_{\geq 0} \times S_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}\) between \(X\) and \(Y\) is defined as:

\[
\lambda\text{-IM-Sig}(X, Y) = \min_{f \in F} \left\{ \frac{1}{\lambda} \sum_{x \in F} \sum_{y \in F} f(x, y) \cdot \delta(x, y) \right\}
\]

subject to

- non-negativity constraint: \(\forall x, y \in F : f(x, y) \geq 0\),
- source constraint: \(\forall x \in F : \sum_{y \in F} f(x, y) \leq X(x)\), and \(\lambda\)-\textsc{IM-Sig}
- target constraint: \(\forall x, y \in F : f(x, y) \leq Y(y)\), and
- total flow constraint: \(\sum_{x \in F} \sum_{y \in F} f(x, y) = \min\{ \sum_{x \in F} X(x), \sum_{y \in F} Y(y) \}\)
**λ-IM-Sig Lower Bound**

- Constraint relaxation with respect to the target constraint
- Greater solution space than that for the EMD
- Adaptable independent normalization factor \( \lambda \)
- The optimal factor \( \lambda = \min \{ \sum_{x \in F} X(x), \sum_{y \in F} Y(y) \} \) (shown in the paper)
λ-IM-Sig Lower Bound

**λ-IM-Sig lower-bounds EMD on reduced signatures**

Given feature signatures \( X, Y \in S^{\geq 0} \) with total weights
\[
m_X = \sum_{x \in F} X(x) \leq \sum_{y \in F} Y(y) = m_Y,
\]
a reduced feature signature \( X' \in S^{\geq 0} \) with respect to \( X \), and \( \lambda \in \mathbb{R}^+ \) with \( \lambda = m_X \), it holds:
\[
\lambda\text{-IM-Sig}(X', Y) \leq EMD(X, Y).
\]
(Proved in the paper)

- \( \lambda\text{-EMD} \) can be defined in a similar way
- As experiments will show later, \( \lambda\text{-IM-Sig} \) lower bound is more efficient than \( \lambda\text{-EMD} \)
- Different signature reduction heuristics possible, such as earth-based (ER) or centroid-based (CR) dimensionality reduction heuristics
Efficiency vs. Dimensionality: $\lambda$-$IM$-$Sig$

- Real world data: ImageNet\[1\]

![Graphs showing efficiency vs. dimensionality for $\lambda$-$IM$-$Sig$.](image)

Efficiency vs. Dimensionality (Imagenet)

- Real world data: ImageNet\textsuperscript{[1]}
- Existing lower bounds on feature signatures: Rubner\textsuperscript{[2]} and IM-Sig\textsuperscript{[3]}

\textbf{Imagenet ; 100,000 data objects ; 100 nn}

- \(\lambda\)-EMD ER 90%
- \(\lambda\)-IM-Sig CR 90%
- \(\lambda\)-IM-Sig CR 90% and IM-Sig
- \(\lambda\)-IM-Sig and \(\lambda\)-EMD-Sig CR 90%
- Rubner and \(\lambda\)-EMD ER 90%
- Rubner and \(\lambda\)-IM-Sig CR 90%
- IM-Sig
- Rubner

Efficiency vs. Dimensionality (Imagenet)

Less query time than for the competitive methods

⇒ The combination of Rubner and $\lambda$-IM-Sig outperforms other combinations or existing methods regarding efficiency and signature size.
Smaller candidate set than for the existing methods
⇒ The combination of Rubner and $\lambda$-IM-Sig shows better selectivity than for existing methods w.r.t signature size
Efficiency vs. Dimensionality (UKBench)

- Real world data: UKBench\(^{[4]}\)

![Graph showing efficiency vs. dimensionality for UKBench data](image)

- Less query time than for the competitive methods
  \(\Rightarrow\) The combination of Rubner and \(\lambda\)-IM-Sig outperforms other combinations or existing methods regarding efficiency and signature size

---

Selectivity vs. Dimensionality (UKBench)

- Real world data: UKBench\[4\]

- Smaller candidate set than for the existing methods
  \[\Rightarrow\] The combination of Rubner and \(\lambda\)-IM-Sig shows better selectivity than for existing methods w.r.t signature size

Conclusion and Outlook

- Feature signature as feature representation model for scientific data
- High computational time complexity of the Earth Mover’s Distance
- Novel lower-bounding filter approximations $\lambda$-\textit{IM-Sig} and $\lambda$-\textit{EMD}
- High efficiency and selectivity providing time cost reduction

- How to extend this work for other domains?
  - e.g. probabilistic data, uncertain data
- Further minimization of the candidate set
- Investigation of the constraint relaxation
Conclusion and Outlook

- Feature signature as feature representation model for scientific data
- High computational time complexity of the Earth Mover’s Distance
- Novel lower-bounding filter approximations $\lambda$-IM-Sig and $\lambda$-EMD
- High efficiency and selectivity providing time cost reduction

- How to extend this work for other domains?
  $\Rightarrow$ e.g. probabilistic data, uncertain data
- Further minimization of the candidate set
- Investigation of the constraint relaxation

Thank you for your attention!
Any questions?