Efficient and Effective Similarity Search over Probabilistic Data based on Earth Mover’s Distance

Jia Xu¹, Zhenjie Zhang², Anthony K.H. Tung², Ge Yu¹

¹Northeastern University, China
²National University of Singapore, Republic of Singapore
Agenda

• Motivation
• Related works
• Definitions
• Index construction
• Algorithms for EMD-based similarity search
• Experimental results
• Conclusion
Motivation

- Probabilistic data management
- Where do probabilities come from?
  - Discrete events?

<table>
<thead>
<tr>
<th>Car ID</th>
<th>Color</th>
<th>Brand</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEU2010</td>
<td>Red</td>
<td>Mazda</td>
<td>0.7</td>
</tr>
<tr>
<td>NEU2010</td>
<td>Silver</td>
<td>Honda</td>
<td>0.3</td>
</tr>
<tr>
<td>SGD2010</td>
<td>Black</td>
<td>BMW</td>
<td>0.5</td>
</tr>
</tbody>
</table>

- Not easy to get the probabilities in real scenario
Motivation

• Probabilistic data management
• Where do probabilities come from?
  – Discrete events?

<table>
<thead>
<tr>
<th>Car ID</th>
<th>Color</th>
<th>Brand</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEU2010</td>
<td>Red</td>
<td>Mazda</td>
<td>0.7</td>
</tr>
<tr>
<td>NEU2010</td>
<td>Silver</td>
<td>Honda</td>
<td>0.3</td>
</tr>
<tr>
<td>SGD2010</td>
<td>Black</td>
<td>BMW</td>
<td>0.5</td>
</tr>
</tbody>
</table>

• Not easy to get the probabilities in real scenario
Motivation

- Where do probabilities come from?
  - Probability histograms are naturally available
Motivation

• Where do probabilities come from?
  – Probability histograms are naturally available

Samples of the pixels

<table>
<thead>
<tr>
<th>Pixel</th>
<th>R</th>
<th>G</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>159</td>
<td>131</td>
<td>52</td>
</tr>
<tr>
<td>P2</td>
<td>28</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>P3</td>
<td>205</td>
<td>113</td>
<td>29</td>
</tr>
<tr>
<td>P4</td>
<td>14</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>
Motivation

- Where do probabilities come from?
  - Probability histograms are naturally available

<table>
<thead>
<tr>
<th>Pixel</th>
<th>R</th>
<th>G</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>159</td>
<td>131</td>
<td>52</td>
</tr>
<tr>
<td>P2</td>
<td>28</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>P3</td>
<td>205</td>
<td>113</td>
<td>29</td>
</tr>
<tr>
<td>P4</td>
<td>14</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>

Motivation

- Where do probabilities come from?
  - Probability histograms are naturally available

Samples of the pixels

<table>
<thead>
<tr>
<th>Pixel</th>
<th>R</th>
<th>G</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>159</td>
<td>131</td>
<td>52</td>
</tr>
<tr>
<td>P2</td>
<td>28</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>P3</td>
<td>205</td>
<td>113</td>
<td>29</td>
</tr>
<tr>
<td>P4</td>
<td>14</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>

- More examples on sensor network, RFID, etc.

Motivation

• Queries on probabilistic histograms
  – Accumulated range query
  – Top-k query

• Similarity search over histograms is more important
Motivation

• Queries on probabilistic histograms
  – Accumulated range query
  – Top-k query

• Similarity search over histograms is more important
Motivation

• Distance metric on probabilistic domain?
  – Euclidean distance?
  – $L_p$ distance?
Motivation

• Distance metric on probabilistic domain?
  – Euclidean distance?
  – $L_p$ distance?
Motivation

• **Distance metric on probabilistic domain?**
  – Euclidean distance?
  – $L_p$ distance?
Motivation

- Distance metric on probabilistic domain?
  - Euclidean distance?
  - $L_p$ distance? Against human intuition

\[ L_1(H_1, H_2) = 2 \]
\[ L_1(H_1, H_3) = 1.15 \]
Motivation

• Intuition of the Earth Mover’s Distance (EMD)
  – Work flow moved between bins
Motivation

• **Intuition of the Earth Mover’s Distance (EMD)**
  – Work flow moved between bins
  – Distance moved between bins

Motivation

• Intuition of the Earth Mover’s Distance (EMD)
  – Work flow moved between bins
  – Distance moved between bins

Motivation

• Intuition of the Earth Mover’s Distance (EMD)
  – Work flow moved between bins
  – Distance moved between bins

Better reflect the human intuition
Motivation

• In this paper, we
  – probabilistic records are represented by histograms
  – discuss the problem of similarity search based on the Earth Mover’s Distance (EMD)
  – present a new indexing scheme to answer similarity queries based on EMD with efficiency and effectiveness
Agenda

• Motivation
• Related works
• Definitions
• Index construction
• Algorithms for EMD-based similarity search
• Experimental results
• Conclusion
Related Works

- **Scan-And-Filter framework**

  ![Diagram of Scan-And-Filter framework]

  - **Preprocessing**: dimensionality reduction
  - **Scan**: the histogram records are scan one by one and filtered by two lower bound filters on EMD
  - **Refinement**: if the lower bound filters cannot prune the record, run complete EMD computation
Related Works

• Our work
  – Does not scan the whole data set
  – Index all records with a forest of B+ trees
  – Easy to be implemented in commercial relational database
  – Achieves high scalability and concurrency
Agenda

• Motivation
• Related works
• Definitions
• Index construction
• Algorithms for EMD-based similarity search
• Experimental results
• Conclusion
Definitions

• Formal definition of EMD
  – Linear programming problem

\[
\text{Minimize:} \quad \frac{\sum_{i,j} f_{ij} \cdot d_{ij}}{\min\{\sum_i p[i], \sum_j q[j]\}}
\]

\[s.t.\]
\[
\forall i : \sum_j f_{ij} = p[i]
\]
\[
\forall j : \sum_i f_{ij} = q[j]
\]
\[
\forall i, j : f_{ij} \geq 0
\]

\[P\]
\[
0.5 \quad 0.2 \quad 0.3
\]

\[Q\]
\[
0.5 \quad 0.2 \quad 0.3
\]

Definitions

• **Formal definition of EMD**
  
  — Linear programming problem

\[
\text{Minimize: } \frac{\sum_{i,j} f_{ij} \cdot d_{ij}}{\min\{\sum_i p[i], \sum_j q[j]\}}
\]

\[
s.t. \quad \forall i : \sum_j f_{ij} = p[i]
\]
\[
\forall j : \sum_i f_{ij} = q[j]
\]
\[
\forall i, j : f_{ij} \geq 0
\]
Definitions

- **Formal definition of EMD**

  - Linear programming problem

\[
\begin{align*}
\text{Minimize:} & \quad \frac{\sum_{i,j} f_{ij} \cdot d_{ij}}{\min \left\{ \sum_i p[i], \sum_j q[j] \right\}} \\
\text{s.t.} & \quad \forall i : \sum_j f_{ij} = p[i] \\
& \quad \forall j : \sum_i f_{ij} = q[j] \\
& \quad \forall i, j : f_{ij} \geq 0
\end{align*}
\]

Definitions

• Formal definition of EMD
  – Linear programming problem

\[
\text{Minimize: } \frac{\sum_{i,j} f_{ij} \cdot d_{ij}}{\min\{\sum_i p[i], \sum_j q[j]\}} \\
\text{s.t.} \\
\forall i: \sum_j f_{ij} = p[i] \\
\forall j: \sum_i f_{ij} = q[j] \\
\forall i, j: f_{ij} \geq 0
\]

Constraint 1:
cannot send more ‘earth’ than there is
Definitions

• **Formal definition of EMD**
  
  – Linear programming problem

\[
\begin{align*}
\text{Minimize:} & \quad \frac{\sum_{i,j} f_{ij} \cdot d_{ij}}{\min\{\sum_i p[i], \sum_j q[j]\}} \\
\text{s.t.} & \quad \forall i : \sum_j f_{ij} = p[i] \\
& \quad \forall j : \sum_i f_{ij} = q[j] \\
& \quad \forall i, j : f_{ij} \geq 0
\end{align*}
\]

**Constraint 1:**
cannot send more ‘earth’ than there is

**Constraint 2:**
cannot receive more ‘earth’ than it can hold

Definitions

- **Formal definition of EMD**
  - Linear programming problem

\[
\text{Minimize: } \frac{\sum_{i,j} f_{ij} \cdot d_{ij}}{\min\{\sum_{i} p[i], \sum_{j} q[j]\}}
\]

\[
s.t. \quad \forall i : \sum_{j} f_{ij} = p[i]
\]
\[
\forall j : \sum_{i} f_{ij} = q[j]
\]
\[
\forall i, j : f_{ij} \geq 0
\]

**Constraint 1:**
cannot send more ‘earth’ than there is

**Constraint 2:**
cannot receive more ‘earth’ than it can hold

**Constraint 3:**
The amount of earth moved should not be a negative number
Definitions

- **Formal definition of EMD**
  - Linear programming problem

\[
\begin{align*}
\text{Minimize:} & \quad \frac{\sum_{i,j} f_{ij} \cdot d_{ij}}{\min\{\sum_i p[i], \sum_j q[j]\}} \\
\text{s.t.} & \quad \forall i : \sum_j f_{ij} = p[i] \\
& \quad \forall j : \sum_i f_{ij} = q[j] \\
& \quad \forall i, j : f_{ij} \geq 0
\end{align*}
\]

- Time complexity: \(O(N^3 \log N)\)

**Constraints**

1. **Constraint 1:**
   - cannot send more ‘earth’ than there is

2. **Constraint 2:**
   - cannot receive more ‘earth’ than it can hold

3. **Constraint 3:**
   - The amount of earth moved should not be a negative number

Agenda

• Motivation
• Related works
• Preliminaries
• Index construction
• Algorithms for EMD-based similarity search
• Experimental results
• Conclusion
Primal-Dual Transformation

**Primal problem (EMD)**

\[
\text{Minimize : } \quad \frac{\sum_{i,j} f_{ij} \cdot d_{ij}}{\min\{\sum_i p[i], \sum_j q[j]\}}
\]

\[\text{s.t.} \quad \forall i : \sum_j f_{ij} = p[i], \quad \forall j : \sum_i f_{ij} = q[j], \quad \forall i, j : f_{ij} \geq 0\]
Primal-Dual Transformation

Primal problem (EMD)

\[ \text{Minimize:} \quad \frac{\sum_{i,j} f_{ij} \cdot d_{ij}}{\min\{\sum_i p[i], \sum_j q[j]\}} \]

\[ \text{s.t.} \quad \forall i : \sum_j f_{ij} = p[i] \]

\[ \forall j : \sum_i f_{ij} = q[j] \]

\[ \forall i, j : f_{ij} \geq 0 \]

Dual problem

\[ \text{Maximize:} \quad \sum_i \phi_i \cdot p[i] + \sum_j \pi_j \cdot q[j] \]

\[ \text{s.t.} \quad \forall i, j : \phi_i + \pi_j \leq d_{ij} \]

\[ \forall i : \phi_i \in \mathbb{R} \]

\[ \forall j : \pi_j \in \mathbb{R} \]
Primal-Dual Transformation

**Primal problem (EMD)**

\[
\text{Minimize :} \quad \frac{\sum_{i,j} f_{ij} \cdot d_{ij}}{\min\{\sum_i p[i], \sum_j q[j]\}}
\]
\[
\text{s.t.} \quad \forall i : \sum_j f_{ij} = p[i]
\]
\[
\forall j : \sum_i f_{ij} = q[j]
\]
\[
\forall i, j : f_{ij} \geq 0
\]

**Dual problem**

\[
\text{Maximize :} \quad \sum_i \phi_i \cdot p[i] + \sum_j \pi_j \cdot q[j]
\]
\[
\text{s.t.} \quad \forall i, j : \phi_i + \pi_j \leq d_{ij}
\]
\[
\forall i : \phi_i \in \mathbb{R}
\]
\[
\forall j : \pi_j \in \mathbb{R}
\]
Primal-Dual Transformation

Primal problem (EMD)

Minimize: \[
\sum_{i,j} f_{ij} \cdot d_{ij} \quad \text{subject to:}
\begin{align*}
\forall i : \sum_j f_{ij} &= p[i] \\
\forall j : \sum_i f_{ij} &= q[j] \\
\forall i, j : f_{ij} &\geq 0
\end{align*}
\]

Dual problem

Maximize: \[
\sum_i \phi_i \cdot p[i] + \sum_j \pi_j \cdot q[j] \quad \text{subject to:}
\begin{align*}
\forall i, j : \phi_i + \pi_j &\leq d_{ij} \\
\forall i : \phi_i &\in \mathbb{R} \\
\forall j : \pi_j &\in \mathbb{R}
\end{align*}
\]
Primal-Dual Transformation

- **Feasible solution**
  - $F = \{f_{ij}\}$ in the primal problem
  - $\Phi = \{\pi_i, \Phi_j\}$ in the dual problem
Primal-Dual Transformation

- **Feasible solution**
  - \( F = \{ f_{ij} \} \) in the primal problem
  - \( \Phi = \{ \pi_i, \Phi_j \} \) in the dual problem

\[
\sum \Phi_j p[i] + \sum \pi_i q[j] \leq \text{EMD}(p,q) \leq \sum f_{ij} d_{ij}
\]
Primal-Dual Transformation

• Feasible solution
  - $F = \{f_{ij}\}$ in the primal problem
  - $\Phi = \{\pi_i, \Phi_j\}$ in the dual problem

$$\sum \Phi_j p[i] + \sum \pi_i q[j] \leq \text{EMD}(p,q) \leq \sum f_{ij}d_{ij}$$

Dual Space
  Lower bound

Primal-Dual Transformation

- Feasible solution
  - $F = \{f_{ij}\}$ in the primal problem
  - $\Phi = \{\pi_i, \Phi_j\}$ in the dual problem

\[
\sum \Phi_j p[i] + \sum \pi_i q[j] \leq \text{EMD}(p, q) \leq \sum f_{ij} d_{ij}
\]

Dual Space
Lower bound

Primal Space (EMD)
Upper bound

Primal-Dual Transformation

\[ \text{Maximize}: \quad \sum_{i} \phi_i \cdot p[i] + \sum_{j} \pi_j \cdot q[j] \]
\[ \text{s.t.} \quad \forall i, j : \phi_i + \pi_j \leq d_{ij} \]
\[ \forall i : \phi_i \in \mathbb{R} \]
\[ \forall j : \pi_j \in \mathbb{R} \]

- Two good properties of dual space
  - Independency
    - constrains are independent to those histograms involved (i.e., \( p \) and \( q \))
    - can derive any feasible solution \( \Phi = \{ \Phi_i, \pi_j \} \) regardless of \( p \) and \( q \)
Primal-Dual Transformation

\[ \begin{align*}
\text{Maximize:} & \quad \sum_i \phi_i \cdot p[i] + \sum_j \pi_j \cdot q[j] \\
\text{s.t.} & \quad \forall i, j : \phi_i + \pi_j \leq d_{ij} \\
& \quad \forall i : \phi_i \in \mathbb{R} \\
& \quad \forall j : \pi_j \in \mathbb{R}
\end{align*} \]

- Two good properties of dual space
  - Independency
    - constrains are independent to those histograms involved (i.e., \( p \) and \( q \))
    - can derive any feasible solution \( \Phi = \{\Phi_i, \pi_j\} \) regardless of \( p \) and \( q \)
  - Lower bounding
    - a feasible solution \( \Phi = \{\Phi_i, \pi_j\} \), derives a lower bound (i.e., \( \sum_i \phi_i \cdot p[i] + \sum_j \pi_j \cdot q[j] \)) for \( \text{EMD}(p,q) \)
    - can use lower bound to filter out those no-hit records
Index Construction

- **Dual space:**
  
  \[
  \text{Maximize:} \quad \sum_i \phi_i \cdot p[i] + \sum_j \pi_j \cdot q[j]
  \]
Index Construction

• **Dual space:**

\[ \text{Maximize: } \sum_i \phi_i \cdot p[i] + \sum_j \pi_j \cdot q[j] \]
Index Construction

• Dual space: 

\[
\begin{align*}
\text{Maximize:} & \quad \sum_i \phi_i \cdot p[i] + \sum_j \pi_j \cdot q[j] \\
\end{align*}
\]
Index Construction

- **Dual space:**

  \[ \text{Maximize: } \sum_i \phi_i \cdot p[i] + \sum_j \pi_j \cdot q[j] \]

  \[ \text{key} = \sum \phi_i p_i \]
Index Construction

• Dual space:

\[ \text{Maximize: } \sum_{i} \phi_{i} \cdot p[i] + \sum_{j} \pi_{j} \cdot q[j] \]

\[ \text{key} = \sum \phi_{i} p_{i} \]
Index Construction

• Dual space:

$$\text{Maximize: } \sum_i \phi_i \cdot p[i] + \sum_j \pi_j \cdot q[j]$$

$$key = \sum \Phi_i p_i$$

$$counterKey = \sum \pi_i q_i$$
Index Construction

- **Dual space:**

  \[
  \text{Maximize:} \quad \sum_{i} \phi_{i} \cdot p[i] + \sum_{j} \pi_{j} \cdot q[j]
  \]

  - \( key = \sum \Phi_{i}p_{i} \)
  - \( \text{counterKey} = \sum \pi_{i}q_{i} \)

  \( key + \text{counterkey} = \sum \Phi_{i}p_{i} + \sum \pi_{i}q_{i} \)
Index Construction

- **Dual space:**

\[
\begin{align*}
\text{Maximize:} & \quad \sum_i \phi_i \cdot p[i] + \sum_j \pi_j \cdot q[j] \\
key &= \sum \phi_i p_i \\
counterKey &= \sum \pi_i q_i \\
key + \text{counterkey} &= \sum \phi_i p_i + \sum \pi_i q_i
\end{align*}
\]

Index Construction

• Dual space:

\[
\begin{align*}
\text{Maximize:} & \quad \sum_i \phi_i \cdot p[i] + \sum_j \pi_j \cdot q[j] \\
key &= \sum \phi_i p_i \\
counterKey &= \sum \pi_i q_i \\
key + \text{counterkey} &= \sum \phi_i p_i + \sum \pi_i q_i
\end{align*}
\]

EMD(P, Q)

Lower bound
Index Construction

- Dual space:

\[
\text{Maximize: } \sum_{i} \phi_{i} \cdot p[i] + \sum_{j} \pi_{j} \cdot q[j]
\]

\[
\text{key} = \sum \phi_{i} \cdot p_{i}
\]

\[
\text{counterKey} = \sum \pi_{i} \cdot q_{i}
\]

\[
\text{key} + \text{counterKey} = \sum \phi_{i} \cdot p_{i} + \sum \pi_{i} \cdot q_{i}
\]
Index Construction

- Dual space:

\[
\text{Maximize: } \sum_i \phi_i \cdot p[i] + \sum_j \pi_j \cdot q[j]
\]

\[
\text{key} = \sum \phi_i \cdot p_i
\]

\[
\text{counterKey} = \sum \pi_i \cdot q_i
\]

\[
\text{key} + \text{counterkey} = \sum \phi_i \cdot p_i + \sum \pi_i \cdot q_i
\]

EMD(P, Q)

Lower bound

Index Construction

- B+ tree filter
Index Construction

• B+ tree filter

\[
key(p, \Phi_l) \in \left[ \min_i (\phi_i + \pi_i) + key(q, \Phi_l) - \theta, \theta - ckey(q, \Phi_l) \right]
\]
Index Construction

- **B+ tree filter**

\[
key(p, \Phi_l) \in \left[ \min_i (\phi_i + \pi_i) + key(q, \Phi_l) - \theta, \theta - ckey(q, \Phi_l) \right]
\]

**Lemma 3.1.** Given a record \( p \) indexed by \( T_l \) and a query record \( q \), it is always valid that

\[
key(p, \Phi_l) \leq EMD(p, q) - ckey(q, \Phi_l)
\]

**Lemma 3.2.**

\[
key(p, \Phi_l) \geq \min_i (\phi_i + \pi_i) + key(q, \Phi_l) - EMD(p, q)
\]
Index Construction

• A forest of B+ trees
Index Construction

• A forest of B+ trees

\[ \Phi_1 = <\Phi_i, \pi_j> \quad \text{and} \quad \Phi_2 = <\Phi_i, \pi_j> \]
Index Construction

- A forest of B+ trees

\[ \Phi_1 = \langle \Phi_i, \pi_j \rangle \]

\[ \Phi_2 = \langle \Phi_i, \pi_j \rangle \]
Agenda

• Motivation
• Related works
• Definitions
• Index Structure Construction
• **Algorithms for EMD-based Similarity Search**
• Experimental results
• Conclusion
Algo. - Range Query

Algo. - Range Query

$T_1$

$T_2$

Algo. - Range Query
Algo. - Range Query

\begin{align*}
T_1 & \quad \begin{array}{c}
S1 \quad S2 \quad S9 \\
S4 \quad S5 \quad S3 \\
S6 \quad S7 \quad S8 \\
\end{array} \\
T_2 & \quad \begin{array}{c}
S15 \quad S9 \\
S3 \quad S5 \\
S14 \quad S12 \quad S7 \\
\end{array} \\
\end{align*}

\[ \text{R-EMD Filter (Sigmod'08)} \]

Algo. - Range Query

\[ T_1 \]

\[ T_2 \]

S1 S2 S9
S4 S5 S3
S6 S7 S8

S15 S9
S3 S5
S14 S12 S7

R-EMD Filter (Sigmod’08)

LB-IM Filter (ICDE’06)

Algo. - Range Query

$T_1$

$T_2$

S3 S5

R-EMD Filter (Sigmod’08)

LB-IM Filter (ICDE’06)

UB_p Filter

Algo. - Range Query

$T_1$  

$T_2$  

S1 S2 S3 S4 S5 S9 S6 S7 S8  

S4 S5 S3 S6 S7 S8 S3 S5 S9  

S15 S9 S3 S5 S14 S12 S7  

S3 S5  

R-EMD Filter (Sigmod’08)  

LB-IM Filter (ICDE’06)  

UB$_p$ Filter  

Exact EMD cal.

Algo. - kNN Query \( (k=2) \)

\[ T_1 \]
\[ T_2 \]

Algo. - kNN Query \((k=2)\)
Algo. - kNN Query \((k=2)\)

\[
T_1 \quad T_2
\]

\[
\text{S1, S2, S3, S4, S5, S6, S7, S8, S15, S9, S13, S5, S14, S12, S7}
\]

\[
\text{CR1} \rightarrow \text{S6, S7, S8} \quad \text{Key}(q, \Phi_1) \quad \text{CR2} \rightarrow \text{S14, S12, S7}
\]

\[
\text{CL1} \rightarrow \text{S5, S4, S3, S2, S1} \quad \text{CL2} \rightarrow \text{S5, S13, S9, S15}
\]

Algo. - kNN Query \( (k=2) \)

**Buffer**

\[
T_1 \quad T_2
\]

\[
S_1, S_2, S_3 \rightarrow S_4, S_5 \rightarrow S_6, S_7, S_8
\]

\[
S_15, S_9 \rightarrow S_{13}, S_5 \rightarrow S_{14}, S_{12}, S_7
\]

CR1 → S6, S7, S8

CL1 → S5, S4, S3, S2, S1

CR2 → S14, S12, S7

CL2 → S5, S13, S9, S15

Algo. - kNN Query (k=2)

$T_1$

$T_2$

CR1

CL1

Buffer

S1 S2 S3 S4 S5 S6 S7 S8

S15 S9

S13 S5

S14 S12 S7

S6 S7 S8

S5 S4 S3 S2 S1

S14 S12 S7

S5 S13 S9 S15

CR2

CL2

S5
Algo. - kNN Query \((k=2)\)
Algo. - kNN Query \((k=2)\)

\[ T_1 \]

\[ T_2 \]

CR1: [S6, S7, S8]
CL1: [S5, S4, S3, S2, S1]

CR2: [S14, S12, S7]
CL2: [S5, S13, S9, S15]

Buffer: [S5]
Algo. - kNN Query \((k=2)\)

\(T_1\)

\(T_2\)

CR1

CL1

Buffer

CR2

CL2

S1 S2 S3 → S4 S5 → S6 S7 S8

S15 S9 → S13 S5 → S14 S12 S7

S6 S7 S8

S5 S4 S3 S2 S1

S14 S12 S7

S5 S13 S9 S15

S5

Buffer

S5 2

S5

S5

S7
Agenda

• Motivation
• Related works
• Definitions
• Index structure construction
• Algorithms for EMD-based similarity search
• Experimental results
• Conclusion
Experimental Results

- **Data set**
  - **RETINA1**
    - Dimension: 96  Cardinality = 3,932
  - **IRMA**
    - Dimension: 199  Cardinality = 10,000
  - **DBLP**
    - Dimension: 8  Cardinality = 250,000
Experimental Results

• Comparison algorithms
  – TBI (Tree-Based-Index)
    • TBI-C: Using *clustering*-based method to find feasible solution
    • TBI-R: Using *random*-sampling-based method to find feasible solution
  – SAR (Scan-And-Refinement)

• Measurement
  – CPU time, # of EMD refinement
Experimental Results: **Range Queries**

**Figure 6:** Effect of threshold on average query CPU time for range queries

**Figure 10:** Effect of data size on range queries
Experimental Results: *k*-NN Queries

![Graphs showing CPU time and EMD refinement vs. database size for k-NN queries.](image)

Figure 12: Effect of *k* on average query CPU time for k-NN queries

Figure 14: Effect of data size for k-NN queries

Conclusions

• We present a new B+ tree-based indexing scheme for the general purposes of similarity search on Earth Mover's Distance

• Our index method relies on the primal-dual theory to construct mapping functions from the original probabilistic space to one-dimensional domain

• Our B+ tree-based index framework is
  – High scalability
  – High efficiency
For more information


**Tree Based Indexing (TBI) System**

TBI is a system focus on indexing multi-dimensional histograms and benefiting the Earth-Mover's Distance-based similarity search. For the B+ tree is used in TBI, our technology can be easily embedded into the real DBMS. The technique is introduced in the following paper:


**Implementation**

Below, the dll of our TBI system and the corresponding help document are available for download. All codes are complied with Microsoft VS 2005 in Windows XP. We have also tested the usability of the dll file under the IDE of Microsoft VS 2008 and Microsoft VS 2010. To understand the core technology of TBI system, please refer to the before-mentioned technical report.

- **TBI Package**  [Windows (zip)]
- **User Guide**  [PDF]
Thank you!
contact: xujia@ise.neu.edu.cn