Core-sets for Fair and Diverse Data Summarization

Sepideh Mahabadi
Microsoft Research

Stojan Trajanovski
Microsoft
Diversity Maximization
Diversity Maximization

Given a set of objects, how to pick a few of them while maximizing diversity?
Applications

- Summarization (e.g. User’s Feed, Video, Documents, Images)
Applications

- **Summarization**
  - User’s feed Generation
    - A set of users
    - Each with a set of messages
      - People who they interact with
      - The channels they are part of
      - ...
    - Which messages to show in their feed?
      - Relevant messages are shown to the users based on user’s likes and replies
      - Need to have diversity in the retrieved summary
Applications

- Summarization (e.g. User’s Feed, Documents, Images)
- Searching
Applications

- Summarization (e.g. User’s Feed, Documents, Images)
- Searching
- Recommendation Systems
  - Movies, News articles
  - Shopping
  - Hiring Candidates e.g. for LinkedIn

Applications

• Summarization (e.g. User’s Feed, Documents, Images)
• Searching
• Recommendation Systems
• ...
Objects
(documents, images, etc)

Feature Vectors

Points in a high dimensional space

Modeling the Objects
The Diversity Maximization problem

**Given:** a set of $n$ points $P$ in a metric space and a parameter $k$,

**Goal:** pick a subset $S \subseteq P$ of $k$ points while maximizing “diversity”.

$k = 3$
Diversity Notions
Diversity I: Minimum Pairwise Distance

**Input:** a set of $n$ vectors $P \subset \mathbb{R}^d$ and a parameter $k \leq d$,

**Goal:** pick $k$ points s.t. the minimum pairwise distance of the picked points is maximized.

$$\min_{p,q \in S} \text{dist}(p, q)$$

- **$O(1)$-approx Greedy Algorithm**
  [RRT’94]
Diversity II: Sum of Pairwise Distances

Input: a set of $n$ vectors $P \subset \mathbb{R}^d$ and a parameter $k \leq d$,

Goal: pick $k$ points s.t. the sum pairwise distances of the picked points is maximized.

$$\sum_{p,q \in S} \text{dist}(p, q)$$

- $O(1)$-approx Local Search Algorithm [HRT’97][AMT’13]
Diversity III: Sum of Nearest Neighbor Distances

**Input:** a set of $n$ vectors $P \subset \mathbb{R}^d$ and a parameter $k \leq d$,

**Goal:** pick $k$ points s.t. the sum of NN distances of the picked points is maximized.

$$
\sum_{p \in S} \min_{q \in S \setminus \{p\}} \text{dist}(p, q)
$$

- Between Min-Pairwise Dist and Sum of Pairwise Dists
- $O(\log k)$-approx Alg [CH’01]
- $O(1)$-approx Alg [BGMS’16]
## Diversity Notions

<table>
<thead>
<tr>
<th>Diversity Notion</th>
<th>Offline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min Pairwise Distance</td>
<td>$\min_{p,q \in S} \text{dist}(p, q)$</td>
</tr>
<tr>
<td>Sum of Pairwise distances</td>
<td>$\sum_{p,q \in S} \text{dist}(p, q)$</td>
</tr>
<tr>
<td>Sum of NN Distances</td>
<td>$\sum_{p \in S} \min_{q \in S \setminus {p}} \text{dist}(p, q)$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Constrained(Fair) Diversity Maximization
Constrained/Fair Diversity Maximization

Input:
• sets of vectors $P_1, \ldots, P_m$, $P = \bigcup_i P_i$
• and $k_1, \ldots, k_m$, $k = \sum_i k_i$
Constrained/Fair Diversity Maximization

Input:
• sets of vectors $P_1, \ldots, P_m$, $P = \bigcup_i P_i$
• and $k_1, \ldots, k_m$, $k = \sum_i k_i$

Goal: pick $k_i$ points $S_i \subset P_i$ s.t. the diversity of the picked points $S = \bigcup_i S_i$ is maximized.

$k_1 = 1$
$k_2 = 2$
Prior Work: Fair Diversity Maximization

<table>
<thead>
<tr>
<th>Diversity Notion</th>
<th>FDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min Pairwise Distance</td>
<td>$\theta(m)$</td>
</tr>
<tr>
<td></td>
<td>[MMM20, AMMM’22]</td>
</tr>
<tr>
<td>Sum of Pairwise distances</td>
<td>$\theta(1)$</td>
</tr>
<tr>
<td></td>
<td>[AMM’13]</td>
</tr>
<tr>
<td>Sum of NN Distances</td>
<td>$\theta(1)$</td>
</tr>
<tr>
<td></td>
<td>[BGMS’16]</td>
</tr>
</tbody>
</table>
Application I: in User’s Feed Generation

• Each message has a posted time
• Goal: show more recent messages and less old ones
• Still need diversity

• Modeling Recency
  • Divide the messages in a month into four groups based on the week they have been posted
  • Set $k_i$ to be higher for more recent weeks

• Data Set: Reddit Messages
  • Messages of a single month (~21000 messages) and divide it into four groups based on the week they appear in
Application II: Movie Recommendation

• Task: Movie recommendation
• Goal: assign budgets for each genre, e.g. comedy, action, drama, ...

• MovieLens Data Set
  • Collection of 4000 movies
  • Group based on the movie genre into 18 groups (e.g. “documentary”, “crime”, “drama”, “action”, ...)
Experimental Results

1. **Need for FDM:** As expected, in the unconstrained version, the recency is not preserved
Experimental Results

1. **Need for FDM**: As expected, in the unconstrained version, the recency is not preserved

2. **Price of Balancedness**: diversity loss by resorting to FDM
   - 1%, for sum-of-pairwise distances
   - 20%, for sum of NN-distances
   - 50%, for minimum pairwise distance
FDM under Big Data Model: Coresets
\[ \text{div}_{k_1, \ldots, k_m} \left( \begin{array}{c} C \end{array} \right) \geq \frac{1}{\alpha} \cdot \text{div}_{k_1, \ldots, k_m} \left( \begin{array}{c} P \end{array} \right) \]
## Theoretical Results

<table>
<thead>
<tr>
<th>Diversity Notion</th>
<th>FDM</th>
<th>Coreset Setting</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min Pairwise Distance</td>
<td>$\theta(m)$</td>
<td>Approx. $O(1)$, Coreset Size $O(k)$ per group</td>
<td>[MMM20]</td>
</tr>
<tr>
<td>Sum of Pairwise distances</td>
<td>$\theta(1)$</td>
<td>(1 + $\epsilon$), Depends on $n$ or aspect ratio</td>
<td>[CPP18]</td>
</tr>
<tr>
<td>Sum of NN Distances</td>
<td>$\theta(1)$</td>
<td>Approx. $O(1)$, Coreset Size $O(k_i^2)$ per group</td>
<td>[This work]</td>
</tr>
</tbody>
</table>

- Algorithms are simple to implement
- Show a new offline algorithm for FDM under Sum-of-NN-Distances
Experimental Results

1. **Need for FDM:** As expected, in the unconstrained version, the recency is not preserved

2. **Price of Balancedness:** diversity loss by resorting to FDM
   - 1%, for sum-of-pairwise distances
   - 20%, for sum of NN-distances
   - 50%, for minimum pairwise distance

3. **Using Coresets**
   - The runtime of the algorithm improves by a factor of $100x$
   - The diversity is only lost by a few percents.
   - No need to recompute the summary of old messages.
Experimental Results

1. **Need for FDM:** As expected, in the unconstrained version, the recency is not preserved

2. **Price of Balancedness:** diversity loss by resorting to FDM
   - 1%, for sum-of-pairwise distances
   - 20%, for sum of NN-distances
   - 50%, for minimum pairwise distance

3. **Using Coresets**
   - The runtime of the algorithm improves by a factor of 100x
   - The diversity is only lost by a few precents.
   - No need to recompute the summary of old messages.

4. **Show superiority of our coreset construction algorithm over Prior work**
## Summary

<table>
<thead>
<tr>
<th>Diversity Notion</th>
<th>FDM</th>
<th>Coreset Setting</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min Pairwise Distance</td>
<td>(\theta(m)) [MMM20, AMMM'22]</td>
<td>(O(1)) (\quad) (O(k)) (\text{per group})</td>
<td>[MMM20]</td>
</tr>
<tr>
<td>Sum of Pairwise distances</td>
<td>(\theta(1)) [AMM13]</td>
<td>((1 + \epsilon)) (\quad) Depends on (n) or aspect ratio</td>
<td>[CPP18]</td>
</tr>
<tr>
<td>Sum of NN Distances</td>
<td>(\theta(1)) [BGMS’16]</td>
<td>(O(m \cdot \log k)) (\quad) (O(k^2)) (\text{per group})</td>
<td>[This work]</td>
</tr>
</tbody>
</table>

- Algorithms are simple to implement
- Showed effectiveness of coresets

Thank you!