AIDEme: An active learning based system for interactive exploration of large datasets
Enhui Huang, Luciano Di Palma, Laurent Cetinsoy, Yanlei Diao, Anna Liu

To cite this version:
Enhui Huang, Luciano Di Palma, Laurent Cetinsoy, Yanlei Diao, Anna Liu. AIDEme: An active learning based system for interactive exploration of large datasets. 2019. hal-02430750

HAL Id: hal-02430750
https://inria.hal.science/hal-02430750
Submitted on 7 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
**Motivation**

- An increasing gap between fast growth of data and limited human ability to comprehend data.
- A growing demand of data analytics tools that can bridge this gap and help the user retrieve high-value content from data more effectively.

**System Overview**

- Consider the data content as a set of records, and the user is interested in some of them but not all.
- In each iteration:
  - the user labels a record as "interesting" or "not interesting",
  - a classification model is built,
  - active learning techniques are employed to select a new record from the unlabeled data source.
- Construct an increasingly-more-accurate model of the user interest.
- Upon convergence, the model is run through the entire data source to retrieve all relevant records.

**Key Techniques**

- **Challenge:** Slow Convergence
- **Novel techniques in AIDEme:**
  1. Factorization
  2. Subspatial convex property: the user interest pattern projected onto a subspace often entails a convex object.
  3. When the subspatial convex property holds, we introduce a Dual-Space Model (DSM).
  4. Optimization using Class Distribution

**Factorized Version Space**

(a) Without factorization

<table>
<thead>
<tr>
<th>Color (B/R)</th>
<th>Size (S/L)</th>
<th>Label</th>
<th>Version Space V (size = 16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B(lack)</td>
<td>L(large)</td>
<td>&gt;, +</td>
<td></td>
</tr>
<tr>
<td>B(lack)</td>
<td>S(mall)</td>
<td>&gt;, +</td>
<td></td>
</tr>
<tr>
<td>R(ed)</td>
<td>L(large)</td>
<td>&gt;, +</td>
<td></td>
</tr>
<tr>
<td>R(ed)</td>
<td>S(mall)</td>
<td>&gt;, +</td>
<td></td>
</tr>
</tbody>
</table>

(b) With factorization

<table>
<thead>
<tr>
<th>Color (B/R)</th>
<th>Size (S/L)</th>
<th>Label</th>
<th>Version Space V (size = 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B(lack)</td>
<td>L(large)</td>
<td>&gt;, +</td>
<td></td>
</tr>
<tr>
<td>B(lack)</td>
<td>S(mall)</td>
<td>&gt;, +</td>
<td></td>
</tr>
<tr>
<td>R(ed)</td>
<td>L(large)</td>
<td>&gt;, +</td>
<td></td>
</tr>
<tr>
<td>R(ed)</td>
<td>S(mall)</td>
<td>&gt;, +</td>
<td></td>
</tr>
</tbody>
</table>

4. Optimization using Class Distribution

**Demonstration**

<table>
<thead>
<tr>
<th>Setup</th>
<th>Iterative Exploration</th>
<th>Final retrieval</th>
<th>Comparison</th>
</tr>
</thead>
</table>

**References**
