User Modeling for Interactive User-Adaptive Collection Structuring

Sebastian Stober, Andreas Nürnberger
Faculty of Computer Science
Otto-von-Guericke University Magdeburg
Contents

- Motivation
- Basics
  - Document Representation
  - (Growing) Self-Organizing Map
  - Generic Adaptation Approach
- Formalization as Quadratic Optimization
- Experiments and Results
- Conclusions
Motivation

- Scenario: exploration of conference proceedings
- Generate an overview map

Better: individual structuring
- … learned from user-interaction with the map (reassigning objects by drag & drop actions)
Seismic-electric effect study of mountain rocks
Measurements of seismic-electric effect (SEE) of mountain rocks in laboratory on guided waves were continued with very wide collection of specially prepared samples ...

preprocessing

stemming filtering
Entropy-based index term selection

seism electr effect study mountain rock measure seism electr effect mountain rock laboratory guide wave collect special prepare sample ...

indexing = counting words/buckets

seism 2, mountain rock 4, wave 1, effect 2, ...

vector = “document fingerprint” (TFxIDF, normalized)
Self-Organizing Map (SOM)

- **Projection** of high-dimensional data vectors to lower dimensional data space (usually 2D) under preservation of neighbourhood relations

![Diagram of SOM](image)
SOM Learning

- competitive learning
- additionally neighborhood relations defined
- all vectors $w_i$ in a neighborhood of the winner neuron $c$ are adjusted:

$$\forall i: w_i = w_i + \nu(c,i) \cdot \delta \cdot (w_i - x(t))$$

- $\nu(i,c)$ : neighborhood function
- $\delta$ : learning rate
Growing SOM
Generic Adaptation Approach

- User manually moves a document
- Similarity measure is adapted
- Other documents are **automatically** assigned to other cells
Generic Adaptation Approach

• Standard similarity measure for documents $x_j$ and $x_k$: inner product:

$$\text{sim}(x_j, x_k) = \sum_{l=1}^{m} x_{jl} \cdot x_{kl}$$

(assuming normalized feature vectors)

• Introduction of feature weights $w_l$ to personalize similarity:

$$\text{sim}(x_j, x_k) = \sum_{l=1}^{m} x_{jl} \cdot w_l \cdot x_{kl}$$

• Initial weights are 1.0

• Weight vector $w$ is used as user model
Generic Adaptation Approach

- Document $d$ assigned to cluster $c_s$:
  \[ \text{sim}(c_s, d) > \text{sim}(c_i, d) \quad \forall i \neq s \]

- Moving $d$ to cluster $c_t$:
  \[ \text{sim}(c_t, d) > \text{sim}(c_i, d) \quad \forall i \neq t \]

- i.e. change weights $w_l$ such that:
  \[
  \sum_{l=1}^{m} x_{jl} \cdot w_l \cdot c_{tl} > \sum_{l=1}^{m} x_{jl} \cdot w_l \cdot c_{sl} \quad \forall s \neq t
  \]

How can these weights be computed?
Problems & Limitations

- So far: heuristics to compute new weights
- No limitations for values of the weights
  - Extreme weighting schemes
- No formal guaranty that all manually moved objects are assigned to their target cell
- No additional constraints (e.g. to increase interpretability)
- New approach: using Quadratic Optimization
Quadratic Optimization

- minimize change of weight vector $w$
  \[ \min_{w \in \mathbb{R}^m} \sum_{l=1}^{m} (w_l - 1)^2 \]

- weights should be non-negative
  \[ w_l \geq 0 \quad \forall 1 \leq l \leq m \]

- sum of the weights should be $m$ (dictionary size)
  \[ \sum_{l=1}^{m} w_l = m \]

- keep all manually moved objects at their position
  \[ \sum_{l=1}^{m} x_{jl} \cdot w_l \cdot c_{tl} > \sum_{l=1}^{m} x_{jl} \cdot w_l \cdot c_{sl} \quad \forall s \neq t \]
Evaluation by User Simulation

modify objects by adding random features
learn map on modified objects

repeat

select an object $o$ to be moved
select most similar cell $c$ for $o$ according to user
move $o$ to $c$

until $o$ could not be moved

<table>
<thead>
<tr>
<th>object selection</th>
<th>cell selection</th>
<th>greedy</th>
<th>random</th>
</tr>
</thead>
<tbody>
<tr>
<td>greedy</td>
<td>greedy</td>
<td>scenario 1</td>
<td>scenario 3</td>
</tr>
<tr>
<td>random</td>
<td>random</td>
<td>scenario 2</td>
<td>scenario 4</td>
</tr>
</tbody>
</table>
Experiment 1 - Setup

- 1914 documents from a scientific news archive represented by 800 index terms
- no class information
- Greedy selection heuristic:
  - Cell with lowest average pairwise (ground truth) similarity
  - Object with lowest average pairwise (ground truth) similarity with all other objects in the cell
- Target cell selection:
  - Cell with highest (ground truth) similarity
Experiment 1 - Results

- Top-10 precision increased to 0.82-0.97 (mean 0.93)
- Moving ~1% of the collection was sufficient
- Random selection did not yield worse results

Simulation terminated too early (system inconsistent)

Many iterations
Experiment 2 - Setup

- 10% (947 documents) from the Banksearch dataset (pre-classified into 4/11 classes) represented by 800 index terms

- Greedy selection heuristic:
  - Cell with highest frequency difference of minority-majority class(es)
  - Object belonging to a minority class with lowest average pairwise (ground truth) sim. with all other objects in the cell

- Target cell selection:
  - Cell with highest (ground truth) similarity having the class of the object to be moved as majority class
Experiment 2 - Results

- Purity, inverse purity and f-measure came close to / exceeded the baseline (due to additional information)
- Top-10 precision decreased after a peak (not optimized by heuristic)
- Manually moving 1-2% of all objects was sufficient
Conclusions

- Proposed and evaluated method for user-adaptive collection structuring based on quadratic optimization
- User model: personalized similarity measure
- Only tested for text – other (non-sparse) data might lead to different performance

Future Work:
- Open problem: Sometimes no solution
- Application to multimedia data
- User study with “real” users
Thank you for your attention!
Experimental Setup

- User study:
  - expensive, time consuming, not objective

- Alternative way: simulate user actions
  - User (ground truth) similarity = initial similarity measure on unmodified objects
Experimental Setup

- User study:
  - expensive, time consuming, not objective
- Alternative way: simulate user actions
  - 2 similarity measures:
  - Select and move object according to a ground truth similarity
  - Measure impact
Growing SOM