User's topological similarity

Serrano, J.\textsuperscript{1}; Díez, F.\textsuperscript{2}; Bellogin, A.\textsuperscript{2}

\textsuperscript{1} UIMP
Santander, España
serranopriego@posgrado.uimp.es

\textsuperscript{2} Information Retrieval Group
Departamento de Ingeniería Informática
Universidad Autónoma de Madrid, Spain
fernando.diez@uam.es, alejandro.bellogin@uam.es
Let’s play

Find the 7 differences
Now ... spot the similarity!
A cup and a doughnut are topologically equivalent!!
Basics

• This paper is about **topology** and **data** or **TDA** (Topological Data Analysis)

• Works on the understanding of the **shape of data**

• Recognizing typical shapes (**patterns**)

• Topology studies shapes that are **invariant** under small deformations
This is a simplicial complex.

This is a simplex.
Betti numbers: characterizes simplicial complexes

- $H_0$: 1
- $H_1$: 1
- $H_2$: 0
\begin{align*}
H_0 &: 1 \\
H_1 &: 1 \\
H_2 &: 0 \\
\end{align*}
Vietoris-Rips Complex computation

Vietoris-Rips complex characterizes the topology of a point set
From users to barcodes

BARCODE BASED SIMILARITIES
From users to barcodes

- Age
- Gender
- Location
- Movie
- Music
  - Director
  - Actor
  - Ratings
  - Genre
  - Ratings
User’s topologies
Barcode Overlapping (Extended) similarity metric

\[ S_{BO}(A, B) = \frac{1}{|A| + |B|} \sup_{a \in A} \frac{a}{b} \sup_{b \in B} \frac{a}{b} + \sup_{a \in A} \frac{a}{b} \sup_{b \in B} \frac{a}{b} \]

\[ S_{BOE} = \begin{cases} 1 & A \neq \emptyset, B \neq \emptyset \\ S_{BO}(A, B) & A = \emptyset, B = \emptyset \\ 0 & \text{any other case} \end{cases} \]
User’s barcode comparisons

$S_{BO}$ similarity computation
Experiments

• Dataset: Movilens 100K
  – 943 users. 1682 items
  – Users represented by means of pairs \((\text{item}, \text{value})\)
  – # user ratings < 50

• Javaplex library
  – To compute invariants as well as similarities among users’ barcodes

• Standard kNN algorithm
Users' profiles simplicial complexes

\[ r = \sqrt{2} \]

\[ r = \sqrt{8} \]
Associated users’ barcodes computed

\[ \text{Dim 0} \]

\[ \text{Dim 1} \]

\[ \sqrt{2} \quad \sqrt{8} \]

\( \rightarrow \) No holes
## Results

<table>
<thead>
<tr>
<th>Similarity</th>
<th>$S_{BOE}$ 50 20</th>
<th>$S_{BOE}$ 100 15</th>
<th>Cosine</th>
<th>Pearson</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>1.1616</td>
<td>1.1482</td>
<td><strong>1.0791</strong></td>
<td>1.5728</td>
<td>1.1217</td>
</tr>
</tbody>
</table>
Conclusions and future work

• Competitive behaviour against standard Cosine or Pearson similarities

• *Cold start* problem

• Accomplish new experiments using different datasets, graph based algorithms, optimization, etc.

• Improve the computation of barcode similarities, highly dependant on dataset size
Merci, Grazie, Gracias, Thank you!