Multidimensional Clustering of MOOC offers

Applying unsupervised learning algorithms FCM and SOM to text descriptions

Kai-Henning Wilker
02.06.2016

Motivation

- Goal: improving of learning & teaching at university
- MOOC = Massive Open Online Course
- MOOCs could be helpful for introduction / review
- Problems:
  - Vast amount of different MOOC offers
  - What search-keywords should be used?

Vision

- Vision: build MOOC recommendation system for students
- Find recommendations using clusters
- Cluster analysis = Grouping objects that are „similar“

Multidimensionional clustering of MOOC offers – Applying unsupervised learning algorithms FCM and SOM to text descriptions

- Clustering only the MOOC text descriptions of English-language MOOCs
- Aim: determining the best clustering method for this task
- Crawling MOOC offers → other project

Project plan

- Goals:
  - Find suitable document representation
  - Detect clusters in reference data set using FCM and SOM
  - Evaluate the quality of the clusters
- Rough timetable:
  - May: research and preliminary look into implementations
  - June: implementing the algorithms and evaluation framework
  - July: evaluating the clusters and documenting the project

Document representation

- Documents (= MOOC descriptions) have to be represented as vectors
- *Bag of words*:
  - Each dimension represents one word
  - Word occurences are counted
  - Normalization by TF-IDF
- Problem: dimension too large
- Possible solutions:
  - Latent Semantic Indexing (LSI)
  - Locality Preserving Indexing (LPI)

Clustering

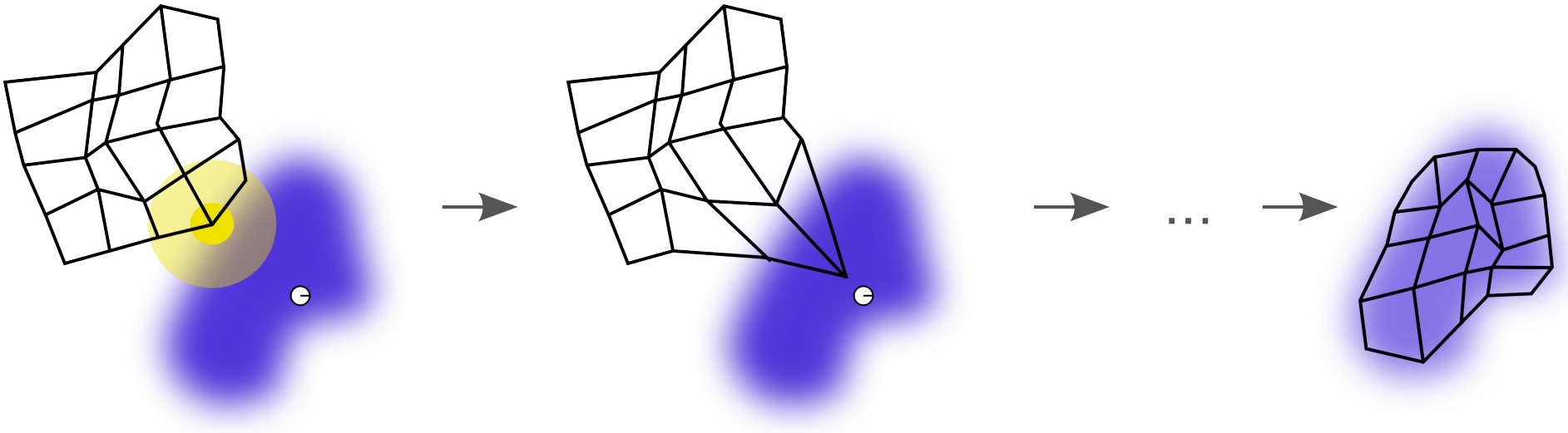
- Many different clustering algorithms available
- Examining *Fuzzy C-Means* and *Self Organisation Maps*
- Both are unsupervised learning algorithms

Fuzzy C-Means

- Derivative of *k-Means* using fuzzy sets
- Documents can belong to more than one cluster → advantage, since MOOCs can have different topics
- Rough description:
  - Initialize *C* cluster centers
  - Repeat:
    - Compute weights of cluster membership for each document
    - Compute new cluster centers

Self Organisation Maps

- SOM is a type of Artifical Neural Network
- Idea: warp a two-dimensional grid onto the set of vectors



Challenges

- Which document representation is the most suitable for each algorithm?
- Determining the values of different parameters which affect the algorithms
  - Number of clusters
  - Fuzzyness parameter
  - Termination criterion
- Labelling the clusters

Evaluation

- External evaluation: let human experts evaluate the quality of the clusters
- Assign validation documents to already existing clusters → compare the results of the algorithms with the classification created by humans
- Additionally, build a simple MOOC search engine using the reference dataset
- Display different clusters with documents relevant to the search query
- Users give feedback whether the clusters are coherent or not

Technical Details

- Manual creation of a test database using MongoDB
- Python as programming language
- Gensim framework for vector representations
- TensorFlow framework for clustering algorithms