Hauptseminar

Recent Developments in Data Science

06.02.2017

Ansprechpartner:
Prof. Dr. Thomas Seidl
Florian Richter

Was ist Data Science?

"**Data science** is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD)."

(Wikipedia: https://en.wikipedia.org/wiki/Data_science)

Ziele

- Einarbeitung in ein wissenschaftliches und aktuelles Thema des Gebiets Data Science
- Auseinandersetzung mit den Fragestellungen und Daten
- Eigene Zusammenfassung mit kreativer Anwendungsidee (~20k – 30k Zeichen)
- Entwicklung eines eigenen durchgehenden Beispiels und exemplarische Anwendung der Methoden (ggfs. durch Implementierung)
- Kritischer Vergleich der eigenen Ergebnisse mit den vorgegebenen Resultaten
- Präsentation der Ergebnisse in einem Vortrag (25min + 10min Fragerunde)

DATABASE SYSTEMS GROUP

Zeitablauf

- 06.02.2017:
  Erstes Treffen, Organisation + Themenvergabe
- 10.02.2017:
  Erste Kontaktaufnahme mit dem Betreuer per Email
- 27.04.2017:
  Start Vortragsserie
  - Donnerstags, 14:00 s.t.
  - Zwei Vorträge pro Treffen
- 29.07.2017 (voraussichtlich):
  Letzter Vortrag
- Zwei Wochen vor Vortrag:
  Abgabe der Ausarbeitung
- Eine Woche vor Vortrag:
  Besprechung des Vortrags und fertige Folien

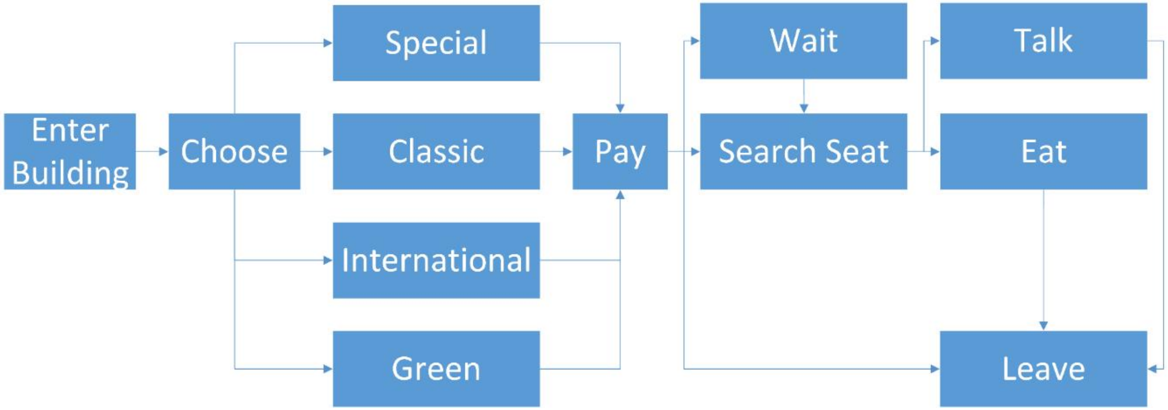
Process Mining

Aktuelle Forschungen

1. Trace Clustering
2. DMN Decision Tables
3. Event Patterns zu Aktivitäten

Mensa Event Log

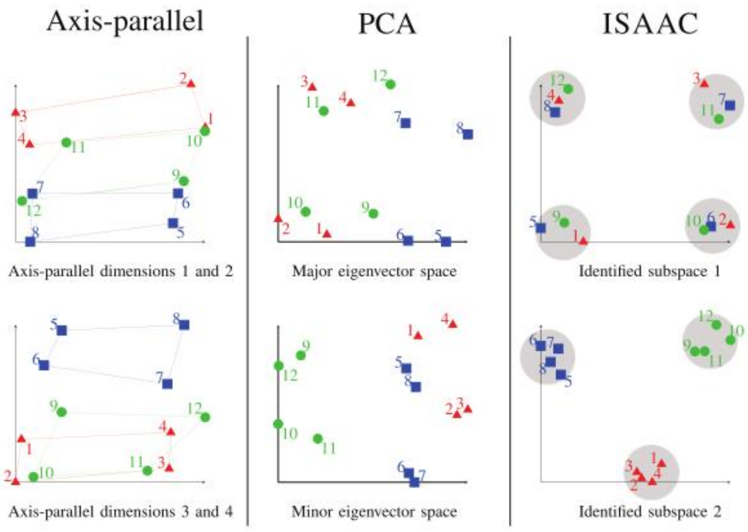
	A	B	C	D	E
1	Student	Activity	Timestamp	GeoData_x	GeoData_y
2	33374094	Enter Building	Wed, 19.10.2016, 13:03	4169	-4555
3	99341331	Pay	Wed, 19.10.2016, 13:05	2393	2144
4	26791916	Pay	Wed, 19.10.2016, 13:06	-3468	1631
5	17879860	Search Seat	Wed, 19.10.2016, 13:09	3067	861
6	20432268	Pay	Wed, 19.10.2016, 13:09	2075	4537
7	87813883	Leave Building	Wed, 19.10.2016, 13:09	-4068	-68
8	71595485	Eat	Wed, 19.10.2016, 13:11	3491	4829
9	32187009	Return Dishes	Wed, 19.10.2016, 13:16	-3515	-3942
10	69178271	Enter Building	Wed, 19.10.2016, 13:20	3104	-2630
11	84309806	Eat	Wed, 19.10.2016, 13:20	1184	1381
12	108009196	Pay	Wed, 19.10.2016, 13:21	4011	793
13	30724430	Eat	Wed, 19.10.2016, 13:23	2219	3979
14	82692348	Leave Building	Wed, 19.10.2016, 13:23	719	-3050
15	27351704	Search Seat	Wed, 19.10.2016, 13:23	-3076	1939
16	101565698	Pay	Wed, 19.10.2016, 13:24	4321	-1655
17	102445419	Eat	Wed, 19.10.2016, 13:28	2669	3023
18	11880894	Eat	Wed, 19.10.2016, 13:30	-3327	4350
19	26299003	Enter Building	Wed, 19.10.2016, 13:31	4874	2256



(Non-linear) clustering

- Ye, W., Maurus, S., Hubig, N., & Plant, C. "Generalized Independent Subspace Clustering" IEEE 16th International Conference on Data Mining, 2016.

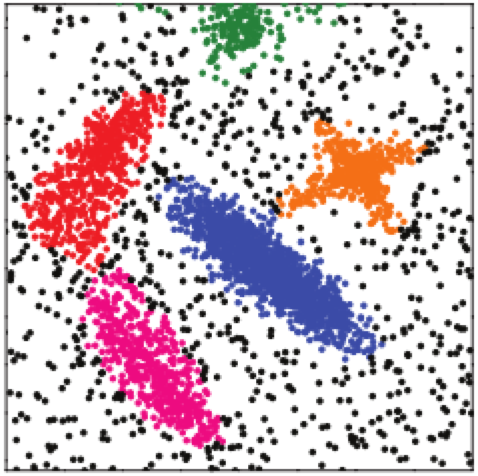
- Subspace clustering
- Minimize statistical dependency between clusters
- Independent Subspace Analysis (ISA)

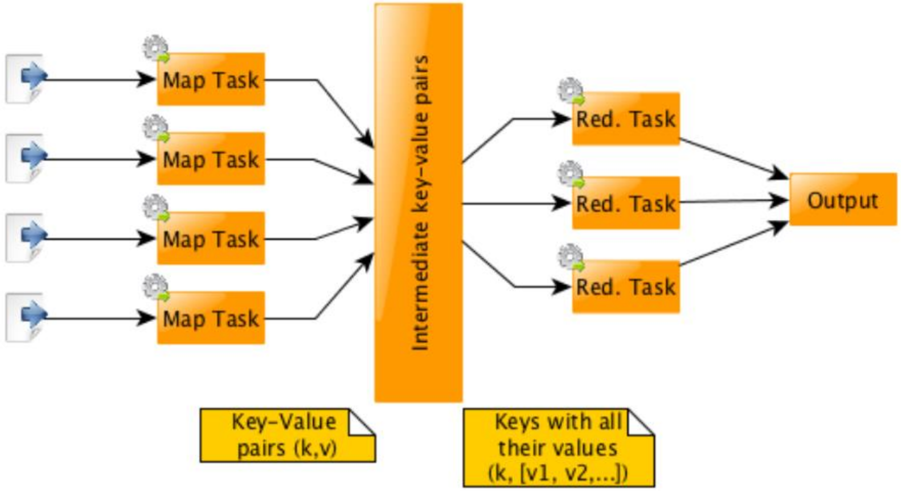


(Non-linear) clustering

- Peng, X.-Y., Yang, Y.-B., Wang, C.-D., Huang, D., & Lai, J.-H. "An Efficient Parallel Nonlinear Clustering Algorithm Using MapReduce" IEEE International Parallel and Distributed Processing Symposium Workshops, 2016.

- MapReduce on non-linear clustering algorithms
- Algorithm used: DenPeak

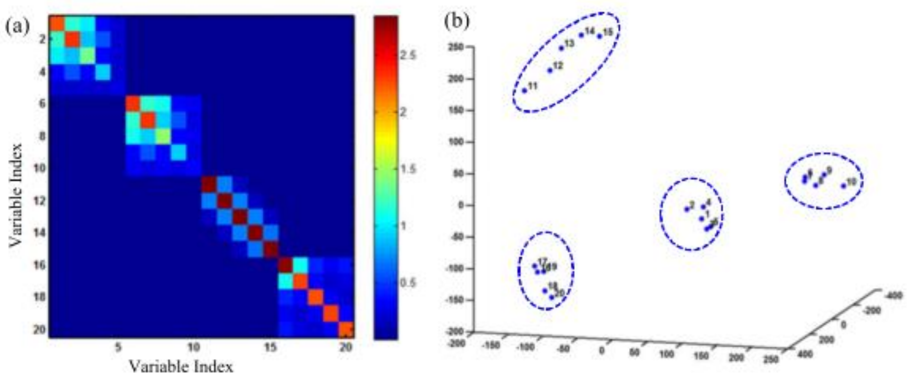




(Non-linear) clustering

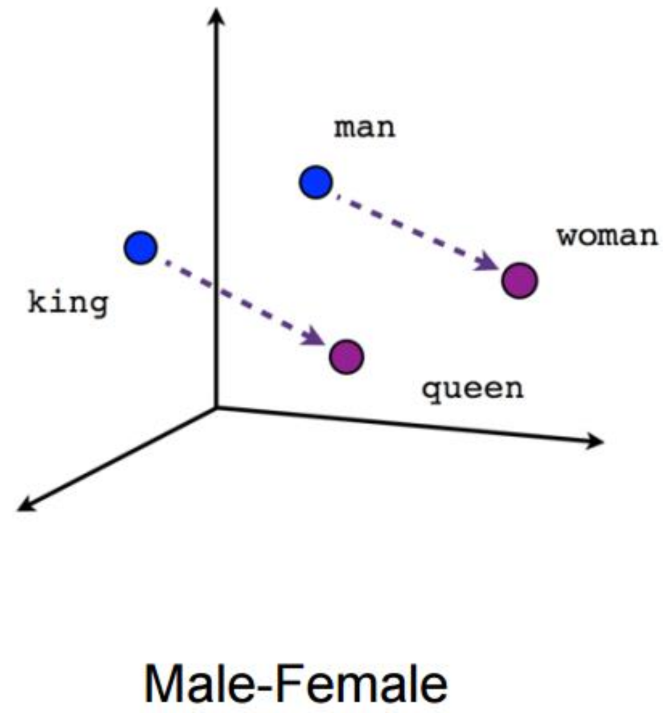
- Chen, Y., & Yang, H. "A Novel Information-Theoretic Approach for Variable Clustering and Predictive Modeling Using Dirichlet Process Mixtures" Nature Scientific Reports, 2016.

- MI for identifying non-linear correlation
- Dirichlet process variable clustering
- Comparing to other MI based non-linear correlation clustering algorithm



Representation Learning

- Representation Learning is an important problem in Machine Learning
- Given: Complex data objects, e.g.
  - Words in a text corpus
  - Nodes in a graph
- Goal: Learn vector representations for these objects, which capture some desired relationships, e.g.
  - Similar semantics
  - Similar neighborhoods in the graph
- The learned embedding allows for applying standard Machine Learning and Data Mining techniques



https://www.tensorflow.org/versions/master/tutorials/word2vec/

Representation Learning

- Word Embedding
  - Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation." EMNLP. Vol. 14. 2014.
- Graph-Node Embedding
  - Cao, Shaosheng, Wei Lu, and Qiongkai Xu. "Grarep: Learning graph representations with global structural information." Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015.
- Semi-Supervised Embedding
  - Weston, Jason, et al. "Deep learning via semi-supervised embedding." Neural Networks: Tricks of the Trade. Springer Berlin Heidelberg, 2012. 639-655.

Representation Learning

- Ordinal Embedding
  - Terada, Yoshikazu, and Ulrike V. Luxburg. "Local ordinal embedding." Proceedings of the 31st International Conference on Machine Learning (ICML-14). 2014.
- Analysis of Event-based Social Networks (EBSN)
  - Pham, Tuan-Anh Nguyen, et al. "A general graph-based model for recommendation in event-based social networks." Data Engineering (ICDE), 2015 IEEE 31st International Conference on. IEEE, 2015.

Prerequisites: Background in / interest to familiarize yourself with Machine Learning and related techniques (in particular Linear Algebra, Probability Theory)

Pattern Mining

- **15:** Finding Subsequences of Patterns
  -> new approach,
  without having to designe a specific encoding scheme

- **16:** Approximating the number of triangles in fully-dynamic streams with fixed memory size
  -> applications: spam/anomaly detection, topic mining

- **17:** Patterns of group evolution
  -> applications: forecasting, pattern discovery

Stream Clustering with Complex Information

- Stream Clustering is a well studied topic.
- Most existing stream clustering algorithms share similar structure: online and offline system.
- The online phrase summaries incoming data points and offline phrase clusters the summarization information.
- However, it is difficult to compute the summarization for data points with complex information and non-euclidean distance.
- In this seminar, we take a look at some state-of-the-art approaches and discuss whether the problem is well addressed.

- **Distance metric learning for soft subspace clustering in composite kernel spacen** (21 Textseiten, viele Formeln)
  - Distance metric learning
    - Die genaue Distanzfunktion wird erst während des Clusterings von den Algorithmen gelernt
  - Soft subspace clustering for high dimensional data
    - Soft (-> Fuzzy) Objekte werden nicht genau einem Cluster voll zugeordnet
    - Subspace clustering: nicht alle Eigenschaften eines Objekts geben zwangsläufig Aufschluss über die Ähnlichkeit zu anderen Objekten
    - High dimensional data: Die Objekte haben viele verschiedene Eigenschaften
  - Composite kernel space
    - Kernel trick: Berechnungen werden implizit in einem höherdimensionalen Raum ausgeführt
- **Pairwise Clustering based on the Mutual-Information Criterion** (27 Textseiten)
  - Pairwise Clustering: Verwendet paarweise Ähnlichkeit zwischen Datenpunkten
  - Based on the Mutual-Information Criterion: Betrachtet die MI zwischen zwei Punkten die hintereinander besucht wurden bei einer Markovkette
- **Network Lasso: Clustering and Optimization in Large Graphs** (21 Textseiten)
  - (Convex optimization, scalability, ADMM (alternating direction method of multipliers)

Themenverteilung

1. From low-level events to activities-a pattern-based approach (2016)
2. Semantics and Analysis of DMN Decision Tables (2016)
3. A non-Compensatory Approach for Trace Clustering (2017)
4. Generalized Independent Subspace Clustering (2016)
5. An Efficient Parallel Nonlinear Clustering Algorithm Using MapReduce (2016)
6. A Novel Information-Theoretic Approach for Variable Clustering and Predictive Modeling Using Dirichlet Process Mixtures (2016)
7. Distance metric learning for soft subspace clustering in composite kernel space (2016)
8. Network lasso: Clustering and optimization in large graphs (2015)
9. Pairwise clustering based on the mutual-information criterion (2016)
10. Grarep: Learning graph representations with global structural information (2015)

Themenverteilung

11. Local ordinal embedding (2014)
12. Deep learning via semi-supervised embedding (2012)
13. A general graph-based model for recommendation in event-based social networks (2015)
14. Glove: Global Vectors for Word Representation (2014)
15. A Subsequence Interleaving Model for Sequential Pattern Mining (2016)
16. TRIÈST: Counting Local and Global Triangles in Fully-Dynamic Streams with Fixed Memory Size (2016)
17. Come-and-Go Patterns of Group Evolution: A Dynamic Model (2016)
18. On graph stream clustering with side information (2013)
19. Stream Clustering: Efficient Kernel-Based Approximation Using Importance Sampling (2015)
20. Streaming spectral clustering (2016)