Hauptseminar

Recent Developments in Data Science

20.10.2016

Ansprechpartner:
Prof. Dr. Thomas Seidl
Florian Richter

Einleitung

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)
- Implementierung einer der vorgestellten Methoden und Anwendung auf eigenes Beispiel
- Kritischer Vergleich der eigenen Ergebnisse mit den vorgegebenen Resultaten
- Präsentation der Ergebnisse in einem Vortrag (35min + 10min Fragerunde)

Zeitablauf

- 20.10.2016:
  Erstes Treffen, Organisation + Themenvergabe
- 24.10.2016:
  Ersten Kontakt mit Betreuer herstellen
- 18.11.2016:
  Spätestens hier erstes Treffen mit Betreuer
  Grobe Gliederung (Sektionsüberschriften) und kurzer Abstract abgeben
- 21.12.2016:
  Bis hier mindestens zweiter Termin mit Betreuer
  Abgabe schriftliche Ausarbeitung (20k-30k Zeichen)
  - Knappe (zusammenfassende) Beschreibung der Methode
  - Ausführliches eigenes Beispiel
- 30.01.2016:
  Bis hier min. dritter Termin mit Betreuer
  Präsentationsfolien fertig
- Mitte Februar:
  Vortragsblock (4x4/5 Vorträge)

Process Mining

Aktuelle Forschungen

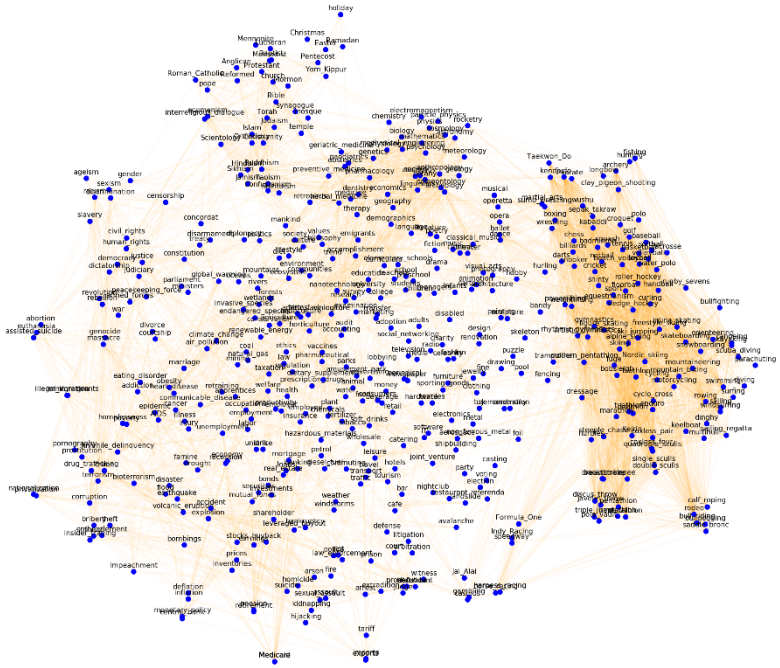
1. Negative Informationen einbetten
2. Ortsinformationen
3. Event Abstraction/ Granularität der Daten
4. Mining von Subprozessen

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

Machine Learning

1. word2vec
   - Mikolov et al.: Efficient Estimation of Word Representations in Vector Space (2013)
   - Mikolov et al.: Distributed Representations of Words and Phrases and their Compositionality (2013)
   - Goal: Learn vector space embeddings of words from huge text corpora
   - The learned embedding captures syntactic and semantic relationships and allows for applying standard machine learning and data mining techniques
   - Easily accessible, several implementations and lots of tutorials available

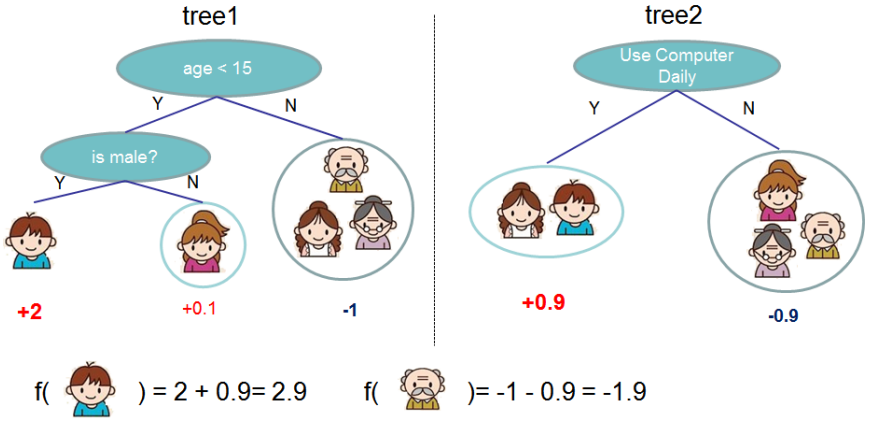


http://www.trivial.io/word2vec-on-databricks/

Machine Learning

2. XGBoost

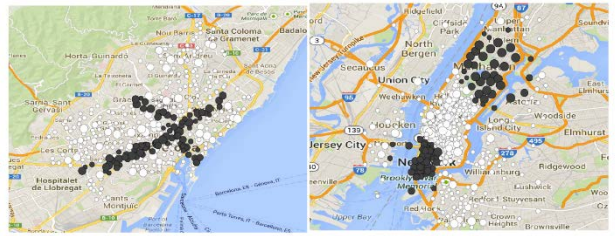
- Chen et al.: XGBoost: A Scalable Tree Boosting System (2016)
- Machine learning method for classification and regression
- Learns an ensemble of regression trees using gradient boosting
- Highly scalable through various extensions
- Widely-used and highly successful in various data science competitions
- Several implementations and lots of tutorials available



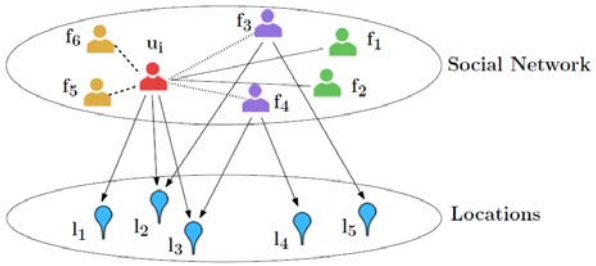
xgboost.readthedocs.io

Geo-Social Analytics

1. Event detection in activity networks
   - Rozenshtein et al.: Event detection in activity networks (2014)
   - Event = subset of nodes in a network (e.g. bike sharing stations) that are close *and* have high activity levels
   - Formulation as hard graph-theoretic problems
   - Approximation using theory of submodular function maximization and semidefinite programming

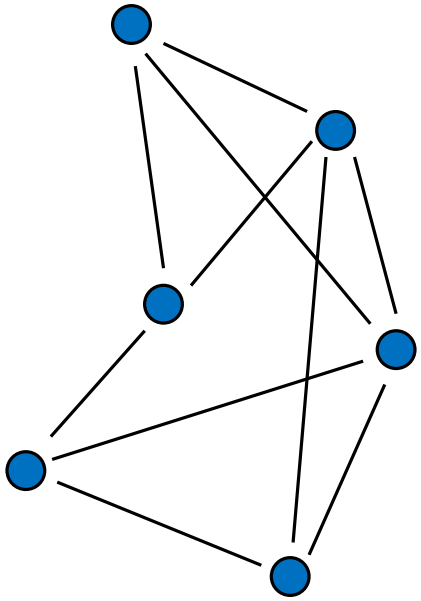


2. Point-of-Interest Recommendations
   - Li et al.: Point-of-Interest Recommendations: Learning Potential Check-ins from Friends (2016)
   - Recommend POIs to users in location-based social networks using information of friends
   - Uses a matrix factorization model



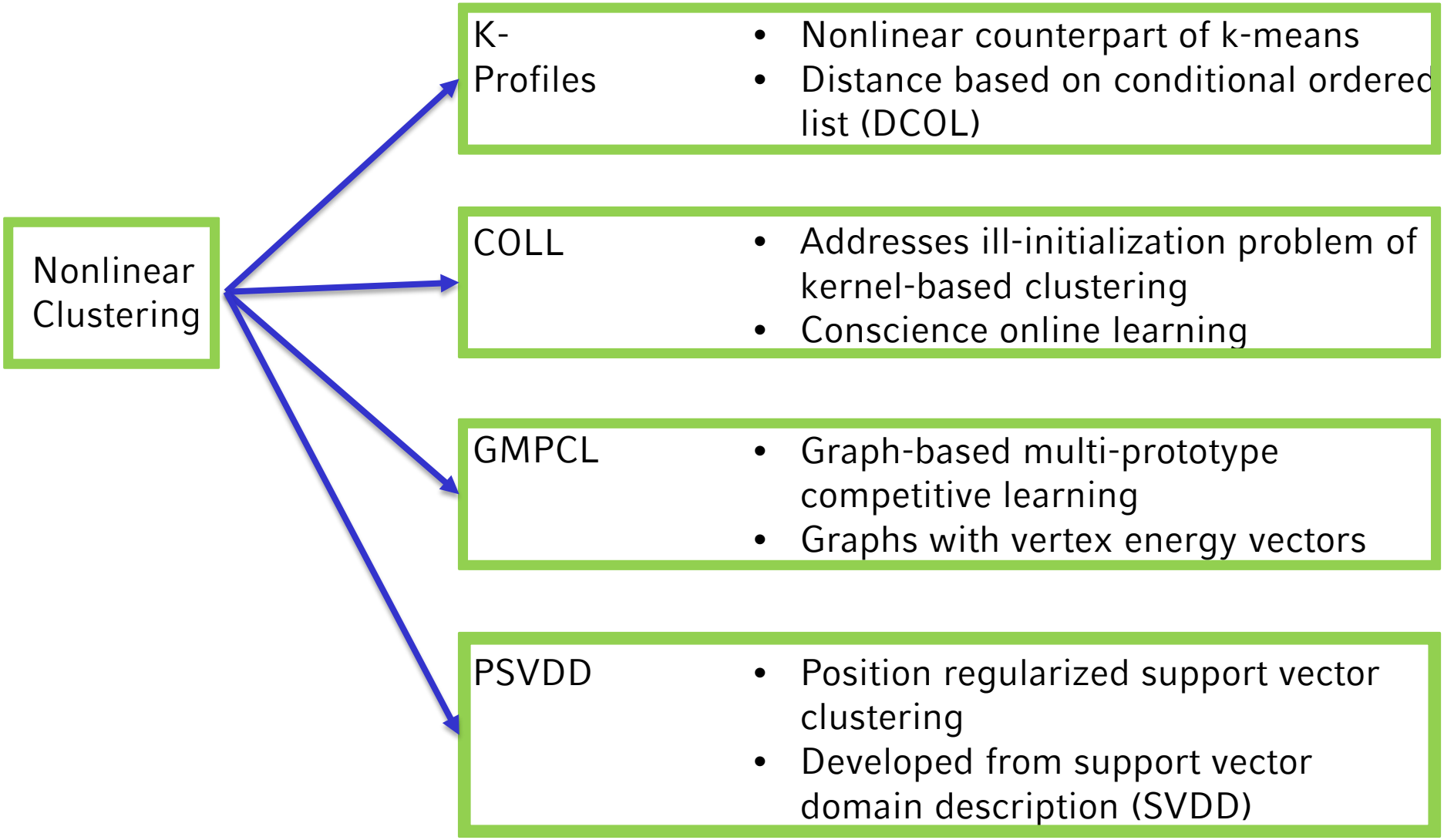
Graph Mining

- Community Search for friend suggestions
- Behaviour between the knodes over time



- Community Search regarding same attributes of the knodes (people)
- Frequent Patterns: optimization in storing them

Nonlinear Clustering



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.

Themenverteilung

1. Incorporating negative information in process discovery (2015)
2. Process Discovery Using Localized Events (2015)
3. Event Abstraction for Process Mining using Supervised Learning Techniques (2016)
4. Mining Local Process Models (2016)
5. Word2vec
6. XGBOOST
7. Event detection in activity networks
8. Point-of-Interest Recommendations: Learning Potential Check-ins from Friends
9. Behavior Query Discovery in System-Generated Temporal Graphs (2015)
10. Ego-net Community Mining Applied to Friend Suggestion (2015)
11. Improved Frequent Pattern Mining Algorithm Based (2015)
12. Effective Community Search for Large Attributed Graphs (2016)
13. K-Profiles A Nonlinear Clustering Method for Pattern
14. Position regularized Support Vector Domain Description
15. Nonlinear Clustering Methods and Applications
16. Graph-based multiprototype competitive learning and its applications
17. Clustering Events on Streams using Complex Context Information
18. On Graph Stream Clustering with Side Information