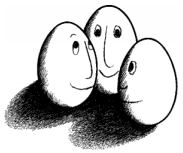
technische universität dortmund



SFB 876 - Providing Information by Resource-Constrained Data Analysis

Proseminar Maschinelles Lernen

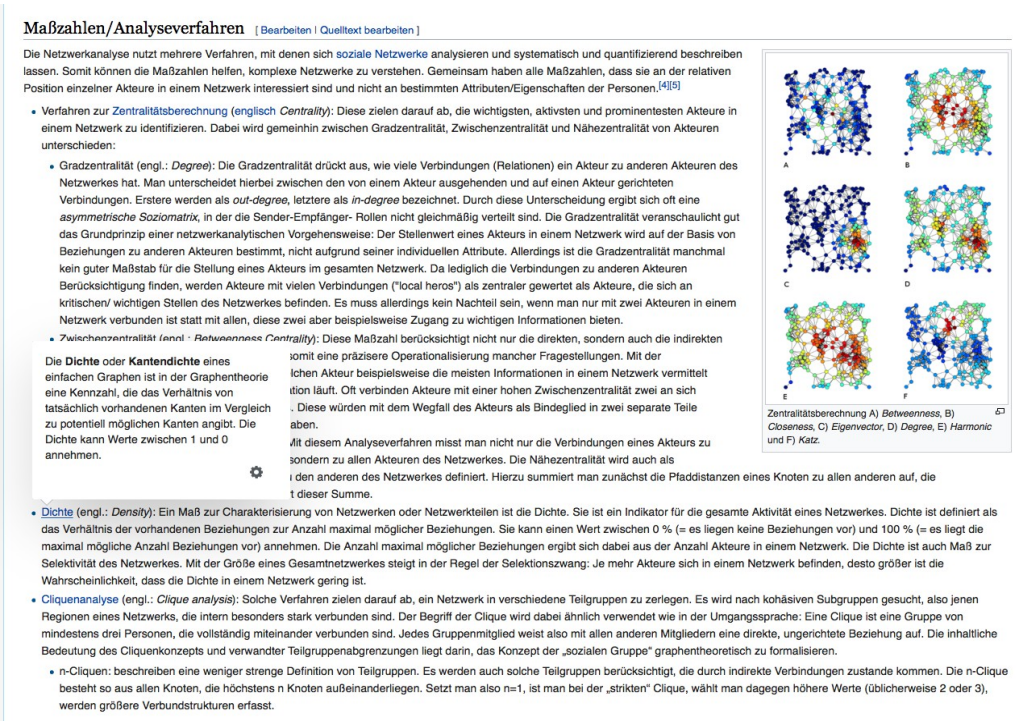
Katharina Morik, Informatik 8, TU Dortmund

Definieren

- Ein X ist ein Oberbegriff (X)
- Gegeben: .... mit...
- Finde: ... so dass

- Zeitschriften
- Konferenzen
- Autoren

- In der Definition verwendete Begriffe müssen ihrerseits definier werden.



Maßzahlen/Analyseverfahren [Bearbeiten | Quelltext bearbeiten]

Die Netzwerkanalyse nutzt mehrere Verfahren, mit denen sich soziale Netzwerke analysieren und systematisch und quantifizierend beschreiben lassen. Somit können die Maßzahlen helfen, komplexe Netzwerke zu verstehen. Gemeinsam haben alle Maßzahlen, dass sie an der relativen Position einzelner Akteure in einem Netzwerk interessiert sind und nicht an bestimmten Attributen/Eigenschaften der Personen.[4][5]

- Verfahren zur Zentralitätsberechnung (englisch *Centrality*): Diese zielen darauf ab, die wichtigsten, aktivsten und prominentesten Akteure in einem Netzwerk zu identifizieren. Dabei wird gemeinhin zwischen Gradzentralität, Zwischenzentralität und Nähezentralität von Akteuren unterschieden:
  - Gradzentralität (engl.: *Degree*): Die Gradzentralität drückt aus, wie viele Verbindungen (Relationen) ein Akteur zu anderen Akteuren des Netzwerkes hat. Man unterscheidet hierbei zwischen den von einem Akteur ausgehenden und auf einen Akteur gerichteten Verbindungen. Erstere werden als *out-degree*, letztere als *in-degree* bezeichnet. Durch diese Unterscheidung ergibt sich oft eine *asymmetrische Soziomatrix*, in der die Sender-Empfänger- Rollen nicht gleichmäßig verteilt sind. Die Gradzentralität veranschaulicht gut das Grundprinzip einer netzwerkanalytischen Vorgehensweise: Der Stellenwert eines Akteurs in einem Netzwerk wird auf der Basis von Beziehungen zu anderen Akteuren bestimmt, nicht aufgrund seiner individuellen Attribute. Allerdings ist die Gradzentralität manchmal kein guter Maßstab für die Stellung eines Akteurs im gesamten Netzwerk. Da lediglich die Verbindungen zu anderen Akteuren Berücksichtigung finden, werden Akteure mit vielen Verbindungen ("local heros") als zentraler gewertet als Akteure, die sich an kritischen/ wichtigen Stellen des Netzwerkes befinden. Es muss allerdings kein Nachteil sein, wenn man nur mit zwei Akteuren in einem Netzwerk verbunden ist statt mit allen, diese zwei aber beispielsweise Zugang zu wichtigen Informationen bieten.
  - Zwischenzentralität (engl.: *Betweenness Centrality*): Diese Maßzahl berücksichtigt nicht nur die direkten, sondern auch die indirekten

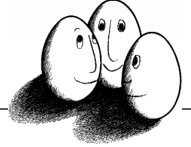
Die **Dichte** oder **Kantendichte** eines einfachen Graphen ist in der Graphentheorie eine Kennzahl, die das Verhältnis von tatsächlich vorhandenen Kanten im Vergleich zu potentiell möglichen Kanten angibt. Die Dichte kann Werte zwischen 1 und 0 annehmen.

omit eine präzisere Operationalisierung mancher Fragestellungen. Mit der lchen Akteur beispielsweise die meisten Informationen in einem Netzwerk vermittelt tion läuft. Oft verbinden Akteure mit einer hohen Zwischenzentralität zwei an sich . Diese würden mit dem Wegfall des Akteurs als Bindeglied in zwei separate Teile aben.

Mit diesem Analyseverfahren misst man nicht nur die Verbindungen eines Akteurs zu sondern zu allen Akteuren des Netzwerkes. Die Nähezentralität wird auch als den anderen des Netzwerkes definiert. Hierzu summiert man zunächst die Pfaddistanzen eines Knoten zu allen anderen auf, die t dieser Summe.

Zentralitätsberechnung A) *Betweenness*, B) *Closeness*, C) *Eigenvector*, D) *Degree*, E) *Harmonic* und F) *Katz*.

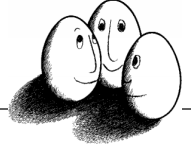
- Dichte (engl.: *Density*): Ein Maß zur Charakterisierung von Netzwerken oder Netzwerkteilen ist die Dichte. Sie ist ein Indikator für die gesamte Aktivität eines Netzwerkes. Dichte ist definiert als das Verhältnis der vorhandenen Beziehungen zur Anzahl maximal möglicher Beziehungen. Sie kann einen Wert zwischen 0 % (= es liegen keine Beziehungen vor) und 100 % (= es liegt die maximal mögliche Anzahl Beziehungen vor) annehmen. Die Anzahl maximal möglicher Beziehungen ergibt sich dabei aus der Anzahl Akteure in einem Netzwerk. Die Dichte ist auch Maß zur Selektivität des Netzwerkes. Mit der Größe eines Gesamtnetzwerkes steigt in der Regel der Selektionszwang: Je mehr Akteure sich in einem Netzwerk befinden, desto größer ist die Wahrscheinlichkeit, dass die Dichte in einem Netzwerk gering ist.
- Cliquenanalyse (engl.: *Clique analysis*): Solche Verfahren zielen darauf ab, ein Netzwerk in verschiedene Teilgruppen zu zerlegen. Es wird nach kohäsiven Subgruppen gesucht, also jenen Regionen eines Netzwerks, die intern besonders stark verbunden sind. Der Begriff der Clique wird dabei ähnlich verwendet wie in der Umgangssprache: Eine Clique ist eine Gruppe von mindestens drei Personen, die vollständig miteinander verbunden sind. Jedes Gruppenmitglied weist also mit allen anderen Mitgliedern eine direkte, ungerichtete Beziehung auf. Die inhaltliche Bedeutung des Cliquenkonzepts und verwandter Teilgruppenabgrenzungen liegt darin, das Konzept der „sozialen Gruppe" graphentheoretisch zu formalisieren.
  - n-Cliquen: beschreiben eine weniger strenge Definition von Teilgruppen. Es werden auch solche Teilgruppen berücksichtigt, die durch indirekte Verbindungen zustande kommen. Die n-Clique besteht so aus allen Knoten, die höchstens n Knoten außeinanderliegen. Setzt man also n=1, ist man bei der „strikten" Clique, wählt man dagegen höhere Werte (üblicherweise 2 oder 3), werden größere Verbundstrukturen erfasst.



Begriffe

a) Soziale Netzwerkanalyse
b) Reinforcement Learning
c) Datenstromalgorithmen
d) Concept drift

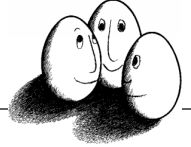
- Unklarer, in verschiedenen Disziplinen gebrauchter Begriff
- Begriff mit einer dominanten Definition
- Begriff mit vielen Spezialisierungen
- Eine Definition kann eine andere Definition einbetten



Soziale Netzwerke, Reinforcement learning

- Soziale Netzwerkanalyse
- Oberbegriff: Soziologie, information retrieval, web search, data mining, massive (big) data
- Gegeben: ? Twitterdaten?
- Finde: communities und ?

- Reinforcement learning
- Oberbegriff: ML, data mining, robotics
- Given: Agents that act, environment that gives a reward
- Find: Sequence of actions in the environment
  - Such that agent receives max. reward
- Autor: Richard Sutton



Datenstromalgorithmen

- Oberbegriff: Algorithmik
- Gegeben: Stream of items, Speicherbeschränkung, Zeit pro Item
- Finde:
- i) frequent items
- ii) set of frequent items
- iii)coresets
- iv)set with maximal diversity
- Autor: Sohler, Cormode

- Nicht nur maschinelles Lernen wird im Datenstrom gemacht.

Concept Drift

- Aus den Daten wird eine Funktion gelernt.
- Die Verteilung der Daten ändert sich.
- Dies muss erkannt werden.
- Dann wird eine neue Funktion für die neue Verteilung gelernt.

Gegeben: Folge von (x_i, y_i)

Finde: $f_\tau(x_i) = \hat{y}_i$

so dass *err* minimal ist

$$err = \sum_{i=1}^{j} |\hat{y}_i - y_i|$$

Erkenne, ob die gelernte Funktion immer noch einen kleinen Fehler hat.

$$i = 1, \ldots j, \ldots n$$

Wenn nein:
Finde
neue Funktion,
So dass *err* minimal ist.

$$f'_\tau(x_j) = \hat{y}'_j$$

$$err = \sum_{j=1}^{n} |\hat{y}'_j - y_j|$$