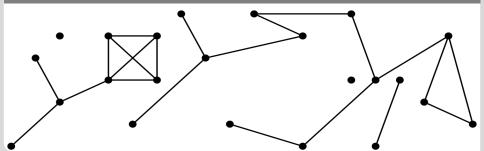


On Node Classification in Dynamic Content-based Networks

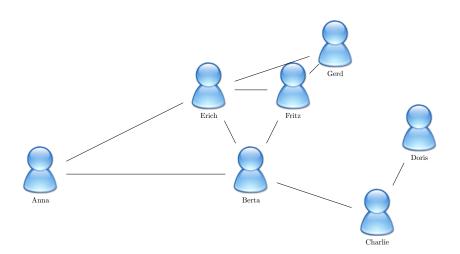
Martin Thoma | 28. Februar 2014

INSTITUT FÜR PROGRAMMSTRUKTUREN UND DATENORGANISATION



Social Network

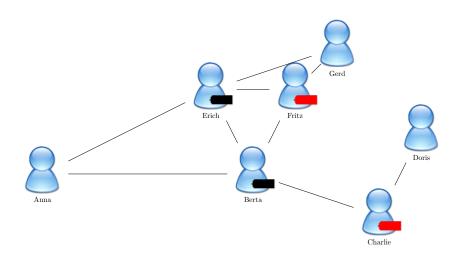




SzenarioÜberblick
000Vokabular
000Sprungtypen
00Zusammenfassung
00Analyse
00Ende
000Martin Thoma − On Node Classification in Dynamic Content-based Networks28. Februar 20142/22

Partially labeled network





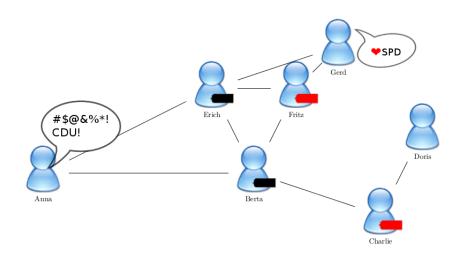
Martin Thoma -	On Node Classifica	ation in Dynamic Co	ntent-based Networks	28. Fel	bruar 2014 3	3/22
0000	0000	000	00	00	00	0000
Szenario	Überblick	Vokabular	Sprungtypen	Zusammenfassung	Analyse	Ende

Partially labeled network with content



Ende

4/22



 Szenario
 Überblick
 Vokabular
 Sprungtypen
 Zusammenfassung
 Analyse

 00●0
 000
 000
 00
 00
 00

 Martin Thoma – On Node Classification in Dynamic Content-based Networks
 28. Februar 2014

Beispiel 2: Literaturdatenbanken



The Development of the C Language Interprocess Communication in the Ninth Edition **Unix System**



Computer Science

The C Programming Language digital restoration and typesetter



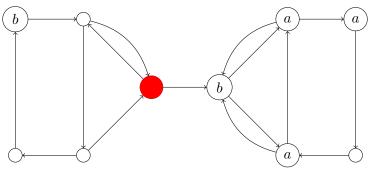
Computer Science

The Identity Thesis for Language and Music



Linguistics





Sprungtypen

Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Überblick

Szenario

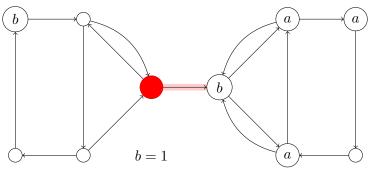
• $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Vokabular

0000	●000	000	00	
Martin Thoma -	On Node	Classification in Dynamic	Content-based	Networks

Analyse





Sprungtypen

Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Überblick

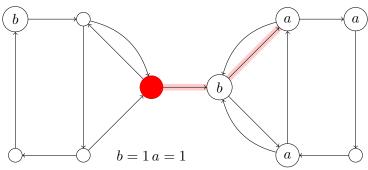
Szenario

■ $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Vokabular

0000	●000	000	00	
Martin Thoma	- On Node Classi	fication in Dynami	c Content-base	d Networks





Sprungtypen

Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Überblick

Szenario

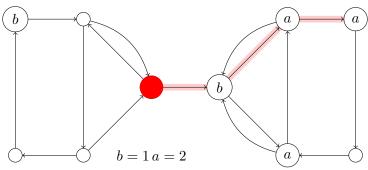
• $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Vokabular

0000	●000	000	00	
Martin Thoma -	On Node	Classification in Dynamic	Content-based	Network

Analyse





Sprungtypen

Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Überblick

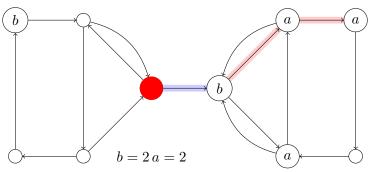
Szenario

• $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$ Vokabular

0000	●000	000	00	
Martin Thoma -	On Node C	lassification in Dynamic	Content-based	Networks

Analyse





Klassifizieren des roten Knotens:

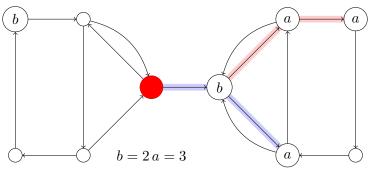
- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Sanaria

• $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$ Vakabular

SECTION	ODCIDICK	VORGDUIGI	Sprungtypen
0000	●000	000	00
Martin Thoma -	On Node Classifica	tion in Dynamic Co	intent-based Netwo





Sprungtypen

Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Überblick

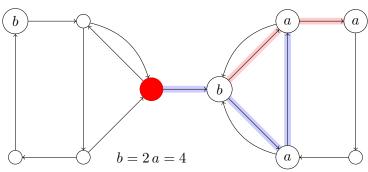
Szenario

• $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$ Vokabular

0000	●000	000	00	0 , 1	
Martin Thoma -	On Node Classifica	ation in Dynamic Co	ontent-	based	Network

Analyse



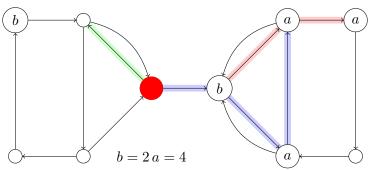


Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $2 \cdot b$ ⇒ Rot mit a klassifizieren

SZCITATIO	ODCIDICK	VORUDUIUI	Sprungtypen
0000	●000	000	00
Martin Thoma -	On Node Classifica	ation in Dynamic Co	intent-based Netwo





Sprungtypen

Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

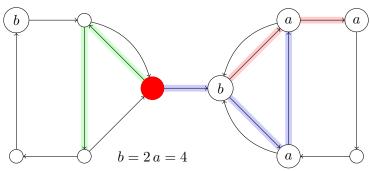
Überblick

Szenario

• $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$ Vokabular

0000	●000	000	00	
Martin Thoma -	On Node	Classification in Dynamic	Content-bas	ed Networks





Sprungtypen

Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Überblick

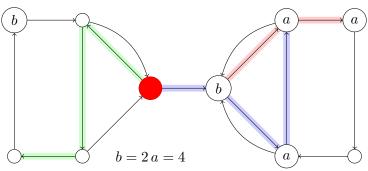
Szenario

• $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$ Vokahular

0000	•000	000	00
Martin Thoma -	On Node Classifica	ition in Dynamic Co	ntent-based Networks

Analyse





Sprungtypen

Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Überblick

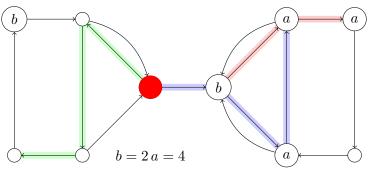
Szenario

■ $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Vokabular

0000	●000	000	00	
Martin Thoma -	On Node	Classification in Dynam	ic Content-based	Networks





Sprungtypen

Klassifizieren des roten Knotens:

- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$ Vakabular

OLCITATIO	OBCIDITOR	v Oltabalai	ob. a 8 c.) be
0000	●000	000	00
Martin Thoma -	On Node Classifica	ation in Dynamic Co	ntent-based Networks

Überblick

Sanaria



- Neben Struktur können Texte genutzt werden



- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!

28. Februar 2014



- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen



- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoter
 - vice versa

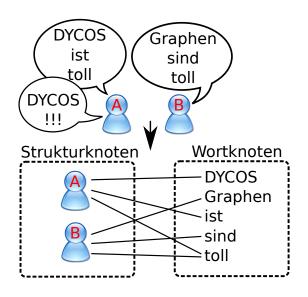


- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoten

Ende

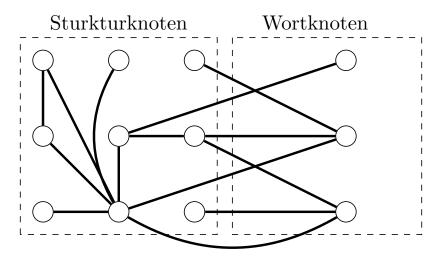


- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoten
 - vice versa



Erweiterter, semi-bipartiter Graph









- Füllwörter: und, oder, im, in, . . .
- ⇒ Beschränkung des Vokabulars sinnvoll

- Zufällige Beispielmenge von Texten für Vokabularbildung betrachten
- Gini-Koeffizient nutzen



- Füllwörter: und, oder, im, in, . . .
- ⇒ Beschränkung des Vokabulars sinnvoll

- Zufällige Beispielmenge von Texten für Vokabularbildung betrachten
- Gini-Koeffizient nutzen



- Füllwörter: und, oder, im, in, . . .
- ⇒ Beschränkung des Vokabulars sinnvoll

- Zufällige Beispielmenge von Texten für Vokabularbildung betrachten
- Gini-Koeffizient nutzen



- Füllwörter: und, oder, im, in, . . .
- ⇒ Beschränkung des Vokabulars sinnvoll

- Zufällige Beispielmenge von Texten für Vokabularbildung betrachten
- Gini-Koeffizient nutzen



- Füllwörter: und, oder, im, in, . . .
- ⇒ Beschränkung des Vokabulars sinnvoll

- Zufällige Beispielmenge von Texten für Vokabularbildung betrachten
- Gini-Koeffizient nutzen



- Füllwörter: und, oder, im, in, . . .
- ⇒ Beschränkung des Vokabulars sinnvoll

- Zufällige Beispielmenge von Texten für Vokabularbildung betrachten
- Gini-Koeffizient nutzen



- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- $g \in (0,1]$
- lack q nahe bei $1 \Rightarrow \text{Wort}$ ist stark ungleich verteilt
- \Rightarrow Nehme Top-m Wörter mit höchstem Gini-Koeffizient

28. Februar 2014



- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- $g \in (0,1]$
- lack g nahe bei $1 \Rightarrow \text{Wort}$ ist stark ungleich verteilt
- \Rightarrow Nehme Top-m Wörter mit höchstem Gini-Koeffizient



- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- $g \in (0,1]$
- $lackbox{ } g$ nahe bei $1 \Rightarrow \text{Wort}$ ist stark ungleich verteilt
- \Rightarrow Nehme Top-m Wörter mit höchstem Gini-Koeffizient



- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- $g \in (0,1]$
- g nahe bei $1 \Rightarrow Wort$ ist stark ungleich verteilt
- \Rightarrow Nehme Top-m Wörter mit höchstem Gini-Koeffizient



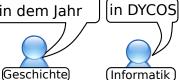
- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- $g \in (0,1]$
- g nahe bei $1 \Rightarrow Wort$ ist stark ungleich verteilt
- \Rightarrow Nehme Top-m Wörter mit höchstem Gini-Koeffizient







in der Schule in dem Jahr



Mathematik

■ Vorkommen insgesamt: 5×

• Vorkommen in "Informatik" $2 \times \Rightarrow p_1 = \frac{2}{\epsilon}$

Vorkommen in "Mathematik" $1 \times \Rightarrow p_2 = \frac{1}{5}$

• Vorkommen in "Geschichte" $2 \times \Rightarrow p_2 = \frac{2}{5}$

• Gini-Koeffizient: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$









in der Schule in dem Jahr



(Geschichte)

- Vorkommen insgesamt: 5×
- Vorkommen in "Informatik" $2 \times \Rightarrow p_1 = \frac{2}{\epsilon}$
- Vorkommen in "Mathematik" $1 \times \Rightarrow p_2 = \frac{1}{5}$
- Vorkommen in "Geschichte" $2 \times \Rightarrow p_2 = \frac{2}{5}$
- Gini-Koeffizient: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$









in der Schule in dem Jahr



Geschichte

- Vorkommen insgesamt: 5×
 - Vorkommen in "Informatik" $2 imes p_1=rac{2}{5}$
 - Vorkommen in "Mathematik" $1 \times \Rightarrow p_2 = \frac{1}{5}$
 - Vorkommen in "Geschichte" $2 \times \Rightarrow p_2 = \frac{2}{5}$
 - Gini-Koeffizient: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$









in der Schule in dem Jahr



Mathematik

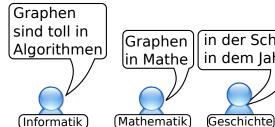
(Geschichte)

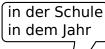
- Vorkommen insgesamt: 5×
 - Vorkommen in "Informatik" $2 \times \Rightarrow p_1 = \frac{2}{5}$

 - Vorkommen in "Geschichte" $2 \times \Rightarrow p_2 = \frac{2}{5}$
 - Gini-Koeffizient: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$











Mathematik Informatik

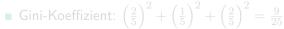
Beispiel: "in"

Vorkommen insgesamt: 5×

• Vorkommen in "Informatik" $2 \times \Rightarrow p_1 = \frac{2}{5}$

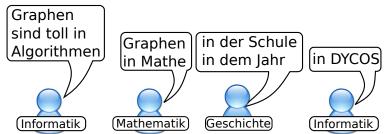
• Vorkommen in "Mathematik" $1 \times \Rightarrow p_2 = \frac{1}{\epsilon}$

• Vorkommen in "Geschichte" $2 \times \Rightarrow p_2 = \frac{2}{5}$









- Vorkommen insgesamt: $5 \times$
- Vorkommen in "Informatik" $2 \times \Rightarrow p_1 = \frac{2}{5}$
- Vorkommen in "Mathematik" $1 \times \Rightarrow p_2 = \frac{1}{5}$
- Vorkommen in "Geschichte" $2 imes\Rightarrow p_2=rac{2}{5}$









in der Schule in dem Jahr



Mathematik)

(Geschichte)

Beispiel: "in"

Vorkommen insgesamt: 5×

• Vorkommen in "Informatik" $2 \times \Rightarrow p_1 = \frac{2}{5}$

• Vorkommen in "Mathematik" $1 \times \Rightarrow p_2 = \frac{1}{5}$

• Vorkommen in "Geschichte" $2 \times \Rightarrow p_2 = \frac{2}{5}$

• Gini-Koeffizient: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 = \frac{9}{25}$

Szenario

Überblick

Vokabular

Sprungtypen

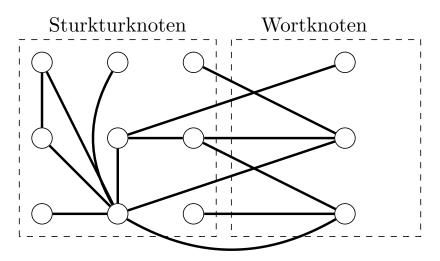
Zusammenfassung

Analyse

Ende

Sprungtypen







Ende

14/22

- ullet Struktursprung: von Strukturknoten v zu Strukturknoten v'
- Inhaltlicher Zweifachsprung: von Strukturknoten v über Wortknoten zu Strukturknoten v^\prime
 - Finde alle Knoten v', die über Wortknoten erreichbar sind (Pfadlänge 2)
 - Nehme Top-*q*-Knoten (Anzahl der Pfade)
 - Wähle zufällig einen davon



- ullet Struktursprung: von Strukturknoten v zu Strukturknoten v'
- **Inhaltlicher Zweifachsprung**: von Strukturknoten v über Wortknoten zu Strukturknoten v'
 - Finde alle Knoten v', die über Wortknoten erreichbar sind (Pfadlänge 2)
 - Nehme Top-*q*-Knoten (Anzahl der Pfade)
 - Wähle zufällig einen davon



- ullet Struktursprung: von Strukturknoten v zu Strukturknoten v'
- Inhaltlicher Zweifachsprung: von Strukturknoten v über Wortknoten zu Strukturknoten v'
 - Finde alle Knoten v', die über Wortknoten erreichbar sind (Pfadlänge 2)
 - Nehme Top-*q*-Knoten (Anzahl der Pfade)
 - Wähle zufällig einen davon



- ullet Struktursprung: von Strukturknoten v zu Strukturknoten v'
- Inhaltlicher Zweifachsprung: von Strukturknoten v über Wortknoten zu Strukturknoten v'
 - Finde alle Knoten v', die über Wortknoten erreichbar sind (Pfadlänge 2)
 - Nehme Top-q-Knoten (Anzahl der Pfade)
 - Wähle zufällig einen davor



- ullet Struktursprung: von Strukturknoten v zu Strukturknoten v'
- Inhaltlicher Zweifachsprung: von Strukturknoten v über Wortknoten zu Strukturknoten v'
 - Finde alle Knoten v', die über Wortknoten erreichbar sind (Pfadlänge 2)
 - Nehme Top-q-Knoten (Anzahl der Pfade)
 - Wähle zufällig einen davon

Wichtige Ideen



- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

Wichtige Ideen



- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

Wichtige Ideen



- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

Dynamisch?



- DYCOS ist nur von der lokalen Situation abhängig
- Klassifizierung von einzelnen Knoten möglich
- Klassifizierung ist einfach

Dynamisch?



- DYCOS ist nur von der lokalen Situation abhängig
- Klassifizierung von einzelnen Knoten möglich
- Klassifizierung ist einfach

28. Februar 2014

Dynamisch?



- DYCOS ist nur von der lokalen Situation abhängig
- Klassifizierung von einzelnen Knoten möglich
- Klassifizierung ist einfach

Datensätze



Alle folgenden Daten sind der Analyse von Aggarwall und Li entnommen.

Name	Knoten	davon beschriftet	Kanten	Beschriftungen
CORA	19 396	14814	75 021	5
DBLP	806 635	18 999	4 414 135	5



Performance:

- Klassifizierung aller Knoten
- Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
- DBLP: < 25 s
- CORA: < 5s
- Klassifikationsgüte:
 - CORA: 82% 84%
 - DBLP: 61% 66%



- Performance:
 - Klassifizierung aller Knoten
 - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
 - DBLP: < 25 s
 - CORA: < 5 s
- Klassifikationsgüte:
 - CORA: 82% 84%
 - DBLP: 61% 66%



- Performance:
 - Klassifizierung aller Knoten
 - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
 - DBLP: < 25 s
 - CORA: < 5 s
- Klassifikationsgüte:
 - CORA: 82% 84%
 - DBLP: 61% 66%



- Performance:
 - Klassifizierung aller Knoten
 - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
 - DBLP: < 25 s
 - CORA: < 5 s
- Klassifikationsgüte:
 - CORA: 82% 84%
 - DBLP: 61% 66%



- Performance:
 - Klassifizierung aller Knoten
 - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
 - DBLP: < 25 s
 - CORA: < 5 s
- Klassifikationsgüte:
 - CORA: 82% 84%
 - DBLP: 61% 66%



- Performance:
 - Klassifizierung aller Knoten
 - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
 - DBLP: < 25 s</p>
 - CORA: < 5 s</p>
- Klassifikationsgüte:
 - CORA: 82% 84%
 - DBLP: 61% 66%



- Performance:
 - Klassifizierung aller Knoten
 - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
 - DBLP: < 25 s
 - CORA: < 5 s</p>
- Klassifikationsgüte:
 - CORA: 82% 84%
 - DBLP: 61% 66%



- Performance:
 - Klassifizierung aller Knoten
 - Intel Xeon 2.5 GHz mit 32 GB RAM, 1 Kern
 - DBLP: < 25 s</p>
 - CORA: < 5 s
- Klassifikationsgüte:
 - CORA: 82% 84%
 - DBLP: 61% 66%

Danke!



Gibt es Fragen?

Martin Thoma - On Node Classification in Dynamic Content-based Networks

Bildquellen



Crystal_Clear_app_personal.png von Wikipedia Commons

Literatur



- Charu C. Aggarwal, Nan Li: On Node Classification in Dynamic Content-based Networks.
- Smriti Bhagat, Graham Cormode und S. Muthukrishnan. Node Classification in Social Networks.
- M. F. Porter. Readings in Information Retrieval. Kapitel An Algorithm for Suffix Stripping.
- Jeffrey S. Vitter. Random Sampling with a Reservoir.

Folien, LaTeXund Material



Der Foliensatz und die LATEX und TikZ-Quellen sind unter github.com/MartinThoma/LaTeXexamples/tree/master/presentations/Datamining-Proseminar Kurz-URL: tinyurl.com/Info-Proseminar

28. Februar 2014

Ende