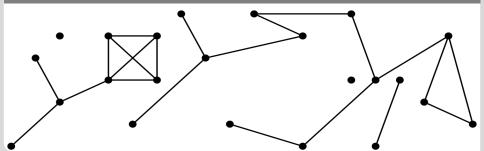


On Node Classification in Dynamic Content-based Networks

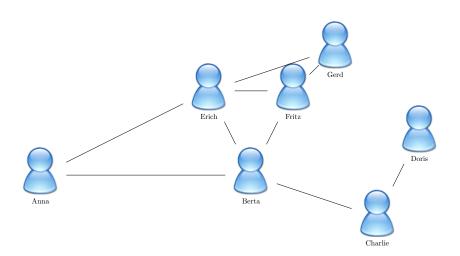
Martin Thoma | 28. Februar 2014

INSTITUT FÜR PROGRAMMSTRUKTUREN UND DATENORGANISATION



Social Network



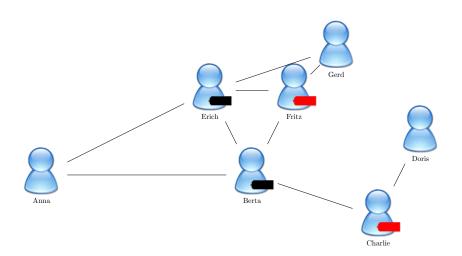


 Szenario
 Überblick
 Vokabular
 Sprungtypen
 Ende

 ●000
 0000
 000
 000
 000

Partially labeled network

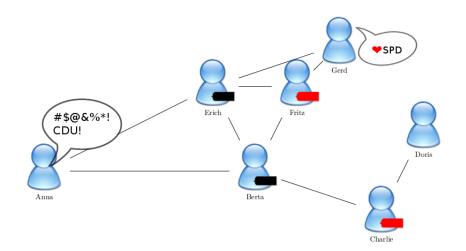




Szenario	Überblick	Vokabular	Sprungtypen	Ende
0000	0000	000	00	0000

Partially labeled network with content





Szenario 00●0 Überblick

Vokabular

Sprungtypen

Ende 000

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

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



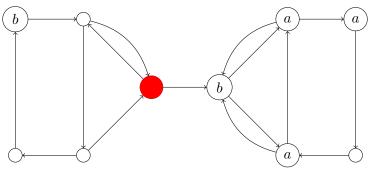
Computer Science

The Identity
Thesis for
Language and
Music



Linguistics





Vokabular

Klassifizieren des roten Knotens:

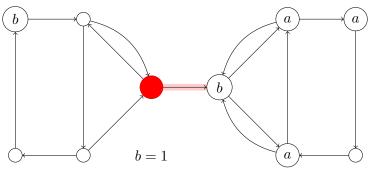
- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Überblick

Szenario

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





Vokabular

Klassifizieren des roten Knotens:

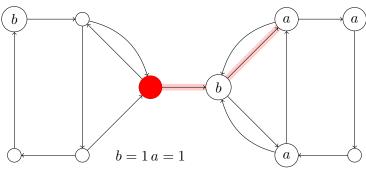
- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Überblick

Szenario

0000		●000		000	
Martin Thoma -	On Node	Classification	in Dynar	nic Content-based	Networks



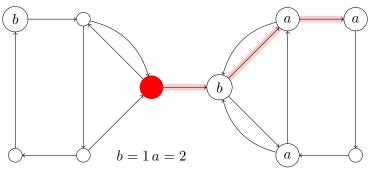


Klassifizieren des roten Knotens:

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

Szenario	Überblick	Vokabular
0000	●000	000
Martin Thoma -	- On Node Classification in Dyna	mic Content based Netwo



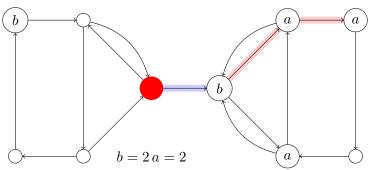


Klassifizieren des roten Knotens:

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

Szenario	Überblick	Vokabular
0000	●000	000
Martin Thoma	On Node Classification in Dyna	mic Content based Net





Vokabular

Klassifizieren des roten Knotens:

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

Überblick

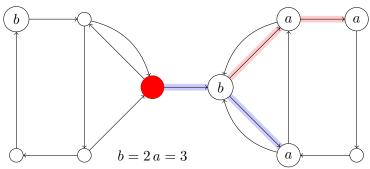
Szenario

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

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

28. Februar 2014



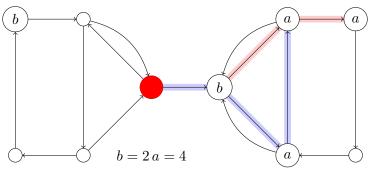


Klassifizieren des roten Knotens:

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

Szenario	Überblick	Vokabular
0000	●000	000
Martin Thoma	- On Node Classification in Dynam	c Content-based Networks





Vokabular

Klassifizieren des roten Knotens:

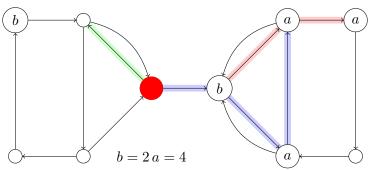
- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Überblick

Szenario

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





Vokabular

Klassifizieren des roten Knotens:

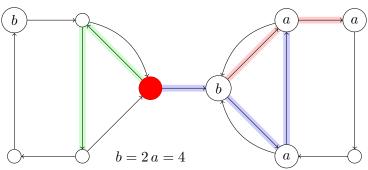
- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Überblick

Szenario

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





Vokabular

Klassifizieren des roten Knotens:

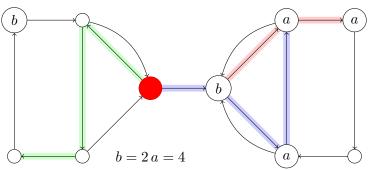
- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Überblick

Szenario

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





Vokabular

Klassifizieren des roten Knotens:

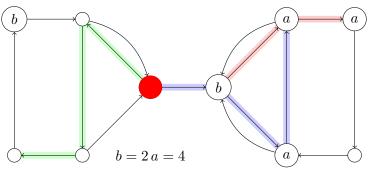
- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk

Überblick

Szenario

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





Klassifizieren des roten Knotens:

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

Überblick

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

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

Szenario



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

28. Februar 2014



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



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



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

28. Februar 2014

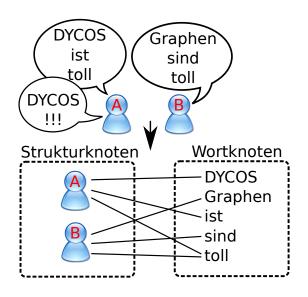


- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoten
 - vice versa

28. Februar 2014

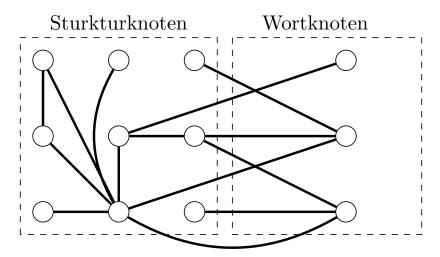


- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoten
 - vice versa



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

Idee:

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

10/18



- 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 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
- $g \in (0,1]$
- $lue{g}$ nahe bei $1 \Rightarrow \text{Wort}$ ist stark ungleich verteilt
- \Rightarrow Nehme Top-m Wörter mit höchstem Gini-Koeffizient

11/18



- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- $g \in (0,1]$
- $lue{g}$ 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



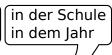
(Geschichte)

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











(Mathematik)

Geschichte

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

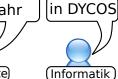








in der Schule in dem Jahr



(Mathematik)

Geschichte

Beispiel: "in"

- Verkemmen incresemt:

• Vorkommen insgesamt: $5 \times$

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

Szenario 0000 Überblick

Vokabular

Sprungtypen

Ende 0000







in der Schule in dem Jahr

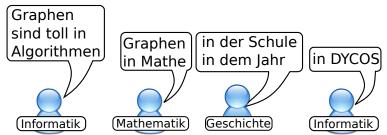


Geschichte

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







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" $2 \times \Rightarrow p_2 = \frac{2}{5}$



28. Februar 2014







in der Schule in dem Jahr



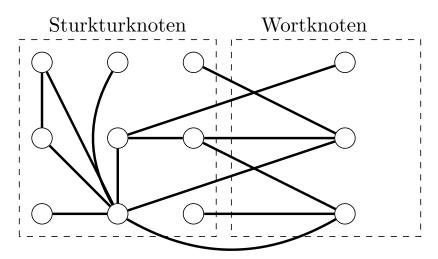
Mathematik)

(Geschichte)

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

Sprungtypen







Überblick

Vokabular

Sprungtypen ●○

28. Februar 2014

Ende 0000



- ullet Struktursprung: von Strukturknoten v zu Strukturknoten v'
- Inhaltlicher Mehrfachsprung: 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 Mehrfachsprung: 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 Mehrfachsprung**: 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 Mehrfachsprung: 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



- $lacktriang{lacktriang}{lacktriang}$: von Strukturknoten v zu Strukturknoten v'
- Inhaltlicher Mehrfachsprung: 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

Danke!



Gibt es Fragen?

28. Februar 2014

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