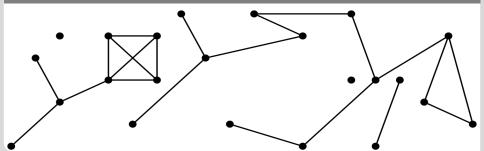


On Node Classification in Dynamic Content-based Networks

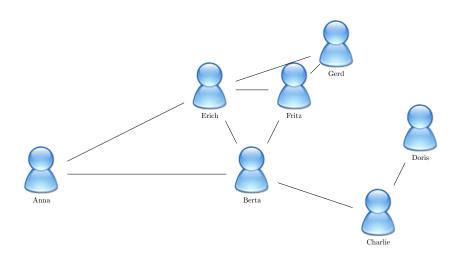
Martin Thoma | 28. Februar 2014

INSTITUT FÜR PROGRAMMSTRUKTUREN UND DATENORGANISATION



Social Network





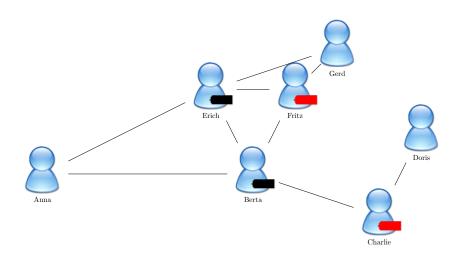
 Szenario
 Überblick
 Vokabular
 Sprungtypen
 Zusammenfassung
 Ende

 ●000
 000
 000
 00
 00
 2

 Martin Thoma – On Node Classification in Dynamic Content-based Networks
 28. Februar 2014
 2/20

Partially labeled network

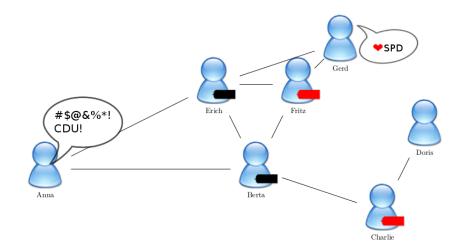




Martin Thoma - On	Node Classification in	Dynamic Content-base	d Networks
0000	0000	000	00
Szenario	Uberblick	Vokabular	Sprungtype

Partially labeled network with content





Szenario 0000

Überblick

Vokabular

Sprungtypen

Zusammenfassung 28. Februar 2014 Ende

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



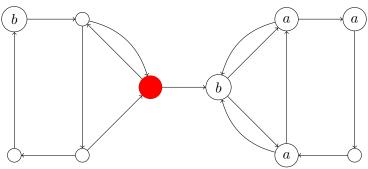
Computer Science

The Identity
Thesis for
Language and
Music



Linguistics

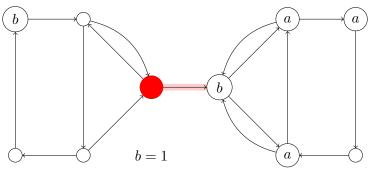




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szei	ıarı	0
000	00	

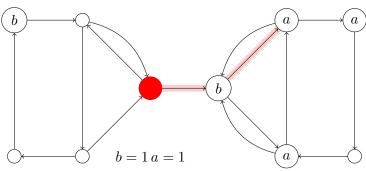




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szei	nari	0
000	00	

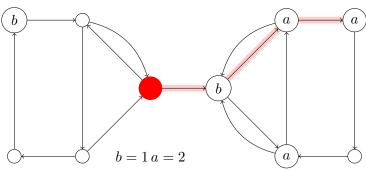




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szei	nari	0
000	00	

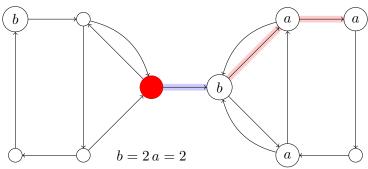




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szei	nari	0
000	00	

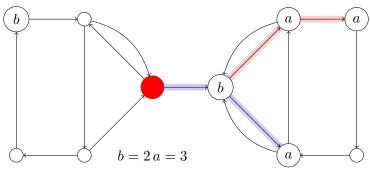




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szei	nari	0
000	00	

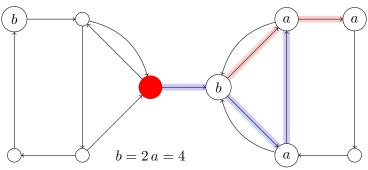




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Sze	nari	0
000	00	

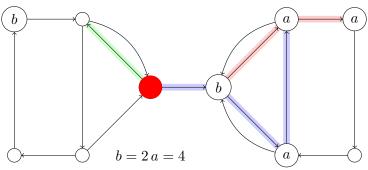




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

10	nari	ze	5	
	00	00	0	

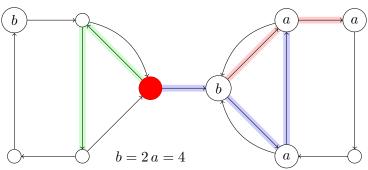




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szena	rio
0000	

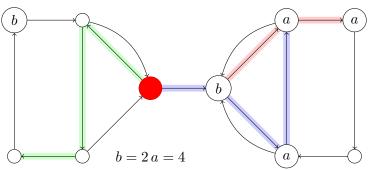




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Szenar	10
0000	

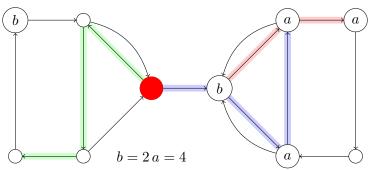




- Zählen von Knotenbeschriftungen in Random Walks
- 4 Random Walks, beginnend bei Rot
- 3 Sprünge pro Random Walk
- $4 \cdot a$, $2 \cdot b \Rightarrow \text{Rot mit } a \text{ klassifizieren}$

Sze	nari	0
000	00	





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

Szenano
0000



- 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 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



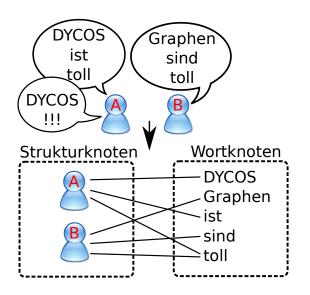
- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoter
 - vice versa



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



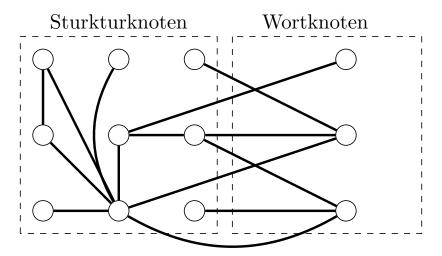
- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoten
 - vice versa



8/20

Erweiterter, semi-bipartiter Graph

















- Füllwörter: und, oder, im, in, ...

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

28. Februar 2014



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



- 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

Geschichte

Beispiel: "in"

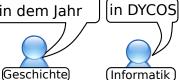
- lacktriangle Vorkommen insgesamt: 5 imes
- Vorkommen in "Informatik" $2 \times \Rightarrow p_1 = \frac{2}{5}$
- Vorkommen in "Mathematik" $1 \times \Rightarrow p_2 = \frac{1}{5}$
- Vorkommen in "Geschichte" $2 \times \Rightarrow p_2 = \frac{2}{5}$
- Gini-Koeffizient: $\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2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







in der Schule in dem Jahr



Beispiel: "in"

■ 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



(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

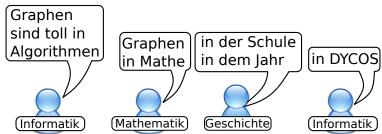


(Mathematik)

Geschichte

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





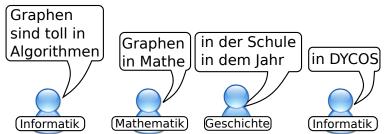
- Vorkommen insgesamt: $5 \times$
- Vorkommen in "Informatik" $2 \times \Rightarrow p_1 = \frac{2}{5}$
- Vorkommen in "Mathematik" $1 \times \Rightarrow p_2 = \frac{1}{5}$
- Vorkommen in "Geschichte" $2 \times \Rightarrow p_2 = \frac{2}{5}$











- 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

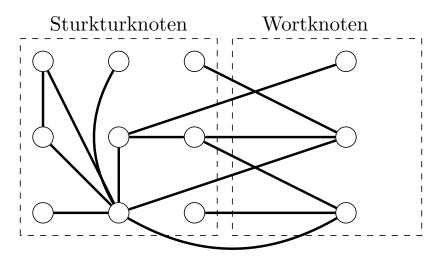


k) Geschichte

- 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

















- 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^\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

Wichtige Ideen



- Random Walk
- Inhaltlicher Zweifachsprung

28. Februar 2014

Ende

Wichtige Ideen



- Random Walk
- Gini-Koeffizient
- Inhaltlicher Zweifachsprung

28. Februar 2014

Ende

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

28. Februar 2014

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

Danke!



Gibt es Fragen?

Ende

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.

28. Februar 2014

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