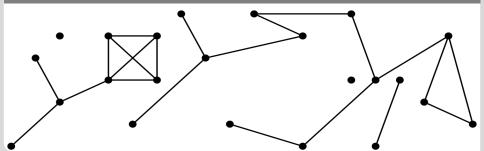


On Node Classification in Dynamic Content-based Networks

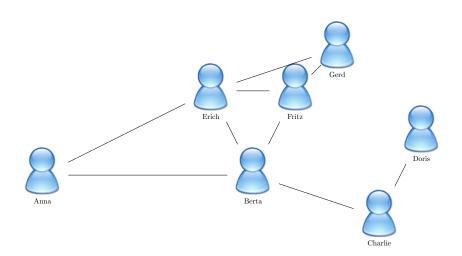
Martin Thoma | 28. Februar 2014

INSTITUT FÜR PROGRAMMSTRUKTUREN UND DATENORGANISATION



Social Network



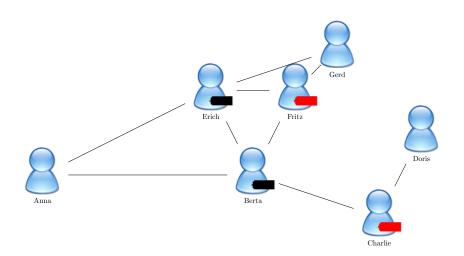


 Szenario
 Überblick
 Vokabular
 Sprungtypen
 Ende

 ●000
 0000
 00
 00
 000

Partially labeled network

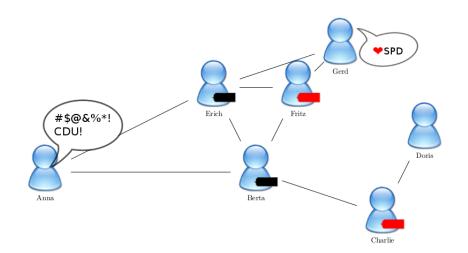




Szenario	Überblick	Vokabular	Sprungtypen	Ende
0000	0000	00	00	0000

Partially labeled network with content





 Szenario
 Überblick
 Vokabular
 Sprungtypen
 Ende

 00●0
 0000
 00
 00
 000

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



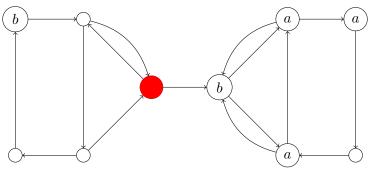
Computer Science

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

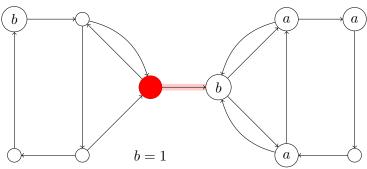
Überblick

Szenario

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

0000	●000	00
Martin Thoma -	On Node Classification in I	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

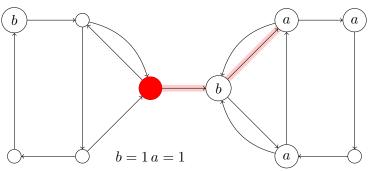
Überblick

Szenario

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

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





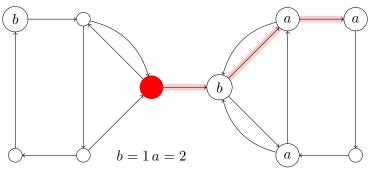
Klassifizieren des roten Knotens:

- 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	00
Martin Thoma	On Mode Classification in Du	namic Content based Notes





Vokabular

Klassifizieren des roten Knotens:

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

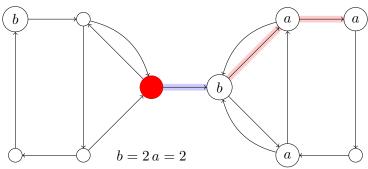
Überblick

Szenario

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

0000	●000	00
Martin 7	Thoma - On Node Classification in Dy	vnamic 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

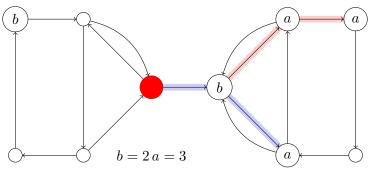
Überblick

Szenario

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

0000	●000	00
Martin Thom	a - On Node Classification in F	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

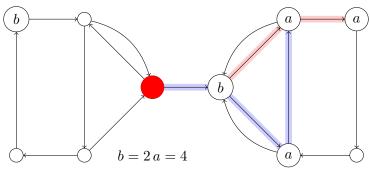
Überblick

Szenario

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

0000	●000	00
Martin Thoma	- On Node Classification in D	Ovnamic 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

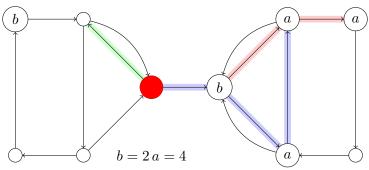
Überblick

Szenario

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

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





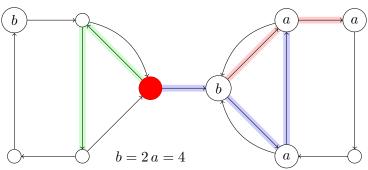
Klassifizieren des roten Knotens:

- 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	00
Martin Thoma -	- On Node Classification in Dynan	nic 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

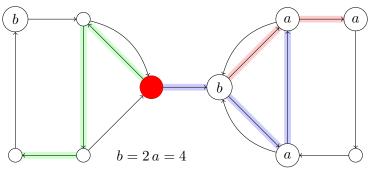
Überblick

Szenario

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

0000	●000	00	
Martin Tho	ma - On Node Classificatio	on in Dynamic Content-bas	sed Networks





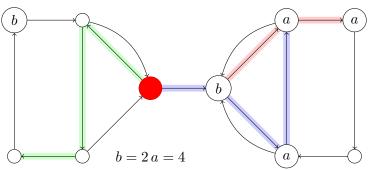
Klassifizieren des roten Knotens:

- 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	00
Martin Thoma On Node	Classification in Dynamic C	ontent based Netwo





Klassifizieren des roten Knotens:

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

Überblick

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

Szenario



- 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



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



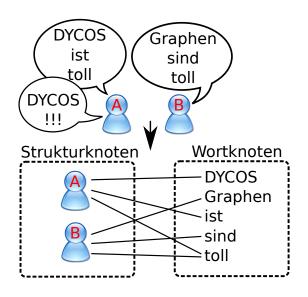
- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoter
 - vice versa



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

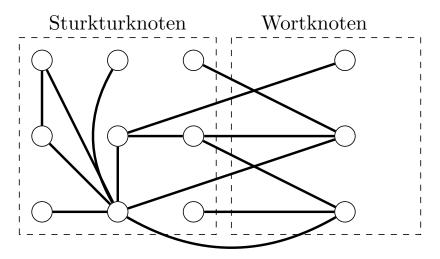


- Neben Struktur können Texte genutzt werden
- Einschränkung: Effizienz!
- Idee: Graph erweitern
 - Texte als Wortmengen
 - Strukturknoten verweisen auf Wortknoten
 - vice versa



Erweiterter, semi-bipartiter Graph







Überblick 000● Vokabular

Sprungtypen

Ende



- Viele Texte ⇒ Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, ...
- ⇒ Beschränkung des Vokabulars sinnvoll

ldee

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



- Viele Texte ⇒ Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, ...
- ⇒ Beschränkung des Vokabulars sinnvoll

Idee

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



- Viele Texte ⇒ Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, ...
- ⇒ Beschränkung des Vokabulars sinnvoll

ldee

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



- Viele Texte ⇒ Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, ...
- ⇒ Beschränkung des Vokabulars sinnvoll

Idee:

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



- Viele Texte ⇒ Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, . . .
- ⇒ Beschränkung des Vokabulars sinnvoll

Idee:

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

10/17



- Viele Texte ⇒ Komplette Textanalyse nicht möglich
- Füllwörter: und, oder, im, in, ...
- ⇒ Beschränkung des Vokabulars sinnvoll

Idee:

- 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
- $g \in (0,1]$
- ullet q 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 $1 \Rightarrow W$ ort ist stark ungleich verteilt
- \Rightarrow Nehme Top-m Wörter mit höchstem Gini-Koeffizient

11/17



- 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-Koeffizient

11/17



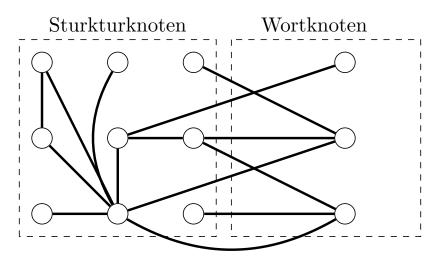
- statistisches Maß für Ungleichverteilung
- $g = \sum_i p_i^2$ mit p_i als relative Häufigkeit
- $g \in (0,1]$
- g nahe bei $1 \Rightarrow Wort$ ist stark ungleich verteilt
- \Rightarrow Nehme Top-m Wörter mit höchstem Gini-Koeffizient



- 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

Sprungtypen









Ende

- $lacktriang{lacktriang}{lacktriang}$: 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



- 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 davor



- $lacktriang{lacktriang}$ 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

Danke!



Gibt es Fragen?

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