Universität Rostock

Traditio et Innovatio

Data- und Graph Mining in Erzähldatenbanken

Zwischenverteidigung der Bachelorarbeit

Tae Keun Jeong
215206849
tae.jeong@uni-rostock.de

Gliederung

1. Einleitung
2. Graph Summarization
3. Aktueller Stand
4. Diskussion
5. Quelle

Einleitung

Motivation

WossiDia/ISEBEL Datensatz : Hypergraph-System

Ein Ansatz, Graphen oder Netzwerk zu verstehen -> Visualisierung

, aber die Visualisierung der großen Graphen mit Millionen Knote

Einleitung

Motivation

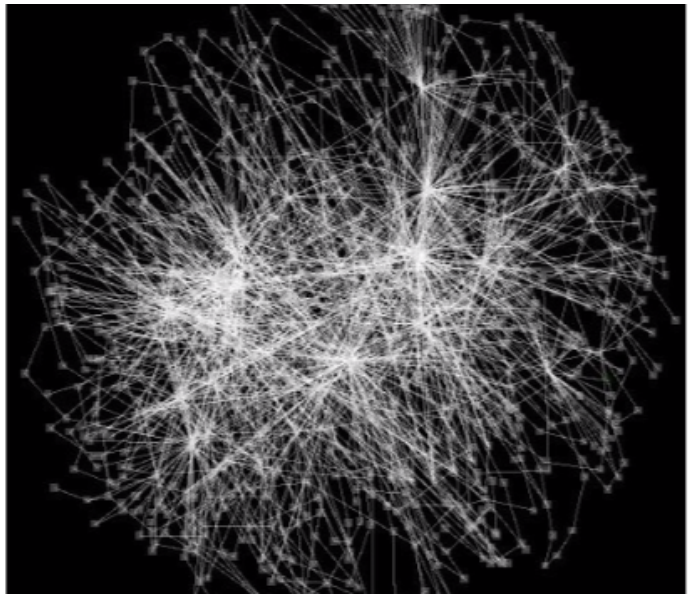


Fig 1. : Ein Graph der Wikipedia-Datensätze

Fig 2 : Haarball

Einleitung

Motivation

1. Ergebnis ist komplex und schwer zu verstehen.
2. Zusammenhänge der Graphenelemente kaum erkennbar

Einleitung

Ziel

1. Zusammenhängende Elemente durch Graphstrukturen zu erkennen.
2. Und dadurch Graphen kompakter darzustellen

Idee : Graph Sumarization

Graph Summarization

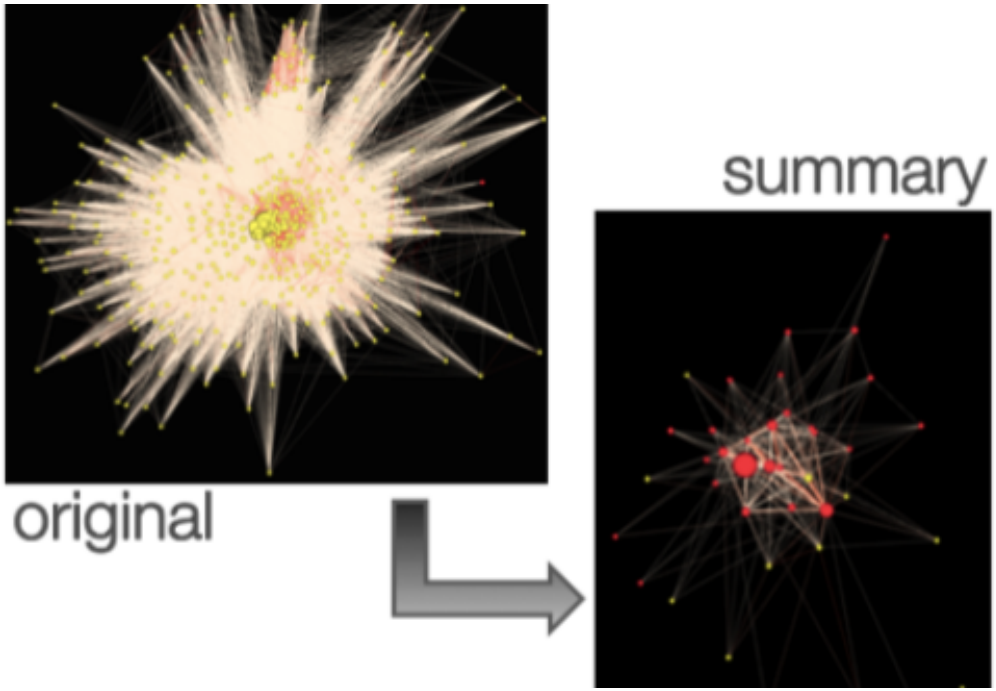


Fig 3. : Beispiel für Graph Summarization

Graph Summarization

Problemdefinition

Gegeben sei ein Graph $G = (V, E), \quad |V| = n, |E| = m$
oder eine $n \times n$ Adjazenzmatrix A von G

Finde eine Menge der Teilgraphen, die möglicherweise überlappend sind und den gegebenen Graphen $G = (V, E)$ so kurz und bündig wie möglich beschreibt

Graph Summarization

Grundlage : Graph

Ein Graph $G = (V, E)$ besteht aus einer Menge von Knoten V(engl. Vertices) und aus einer Menge von Kanten E, $E \subseteq V \times V$ (engl. Edge)

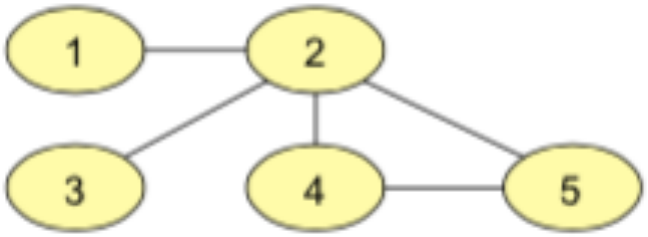
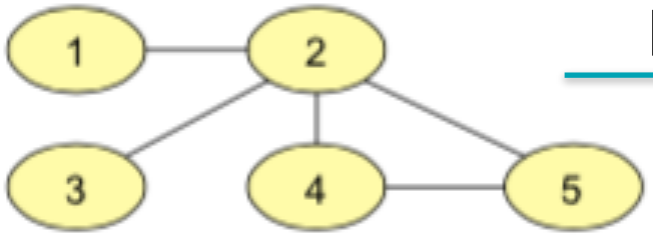


Fig 4 : ein Graph mit |V| = 5 und |E| = 5

Graph Summarization

Grundlage : Adjazenzmatrix A von G

ein Graph $G = (V, E) \quad |V| = n, |E| = m$
eine $n \times n$ Adjazenzmatrix A ist



Kodierung

	1	2	3	4	5
1	0	1	0	0	0
2	1	0	1	1	1
3	0	1	0	0	0
4	0	1	0	0	1
5	0	1	0	1	0

Fig 4 : ein Graph mit IVI = 5 und IEI = 5

Fig 5. : Eine Adjazenzmatrix von Fig 4

0,1 -> Bit

Graph Summarization

Grundlage : Strukturtyp

Fig 6 : (full)Clique

Fig 7 : near-Clique

Graph Summarization

Grundlage : Strukturtyp

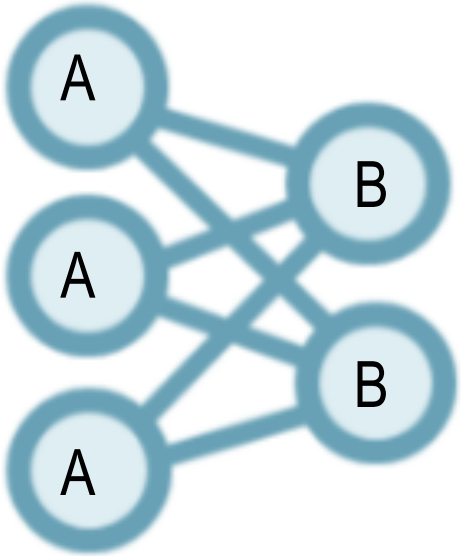


Fig 8 : full-Bipartite

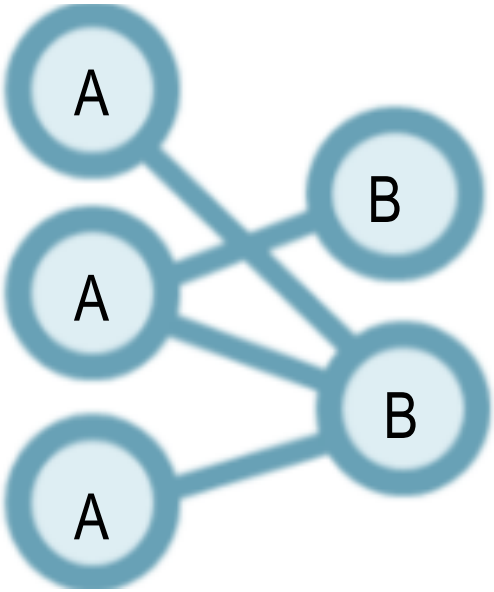


Fig 9 : near-Bipartite

Graph Summarization

Grundlage : Strukturtyp

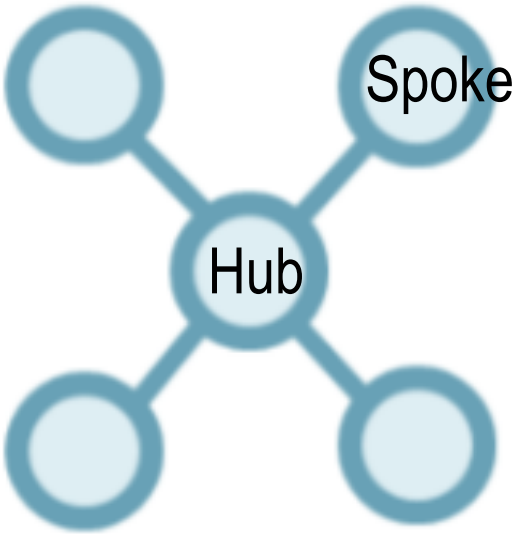
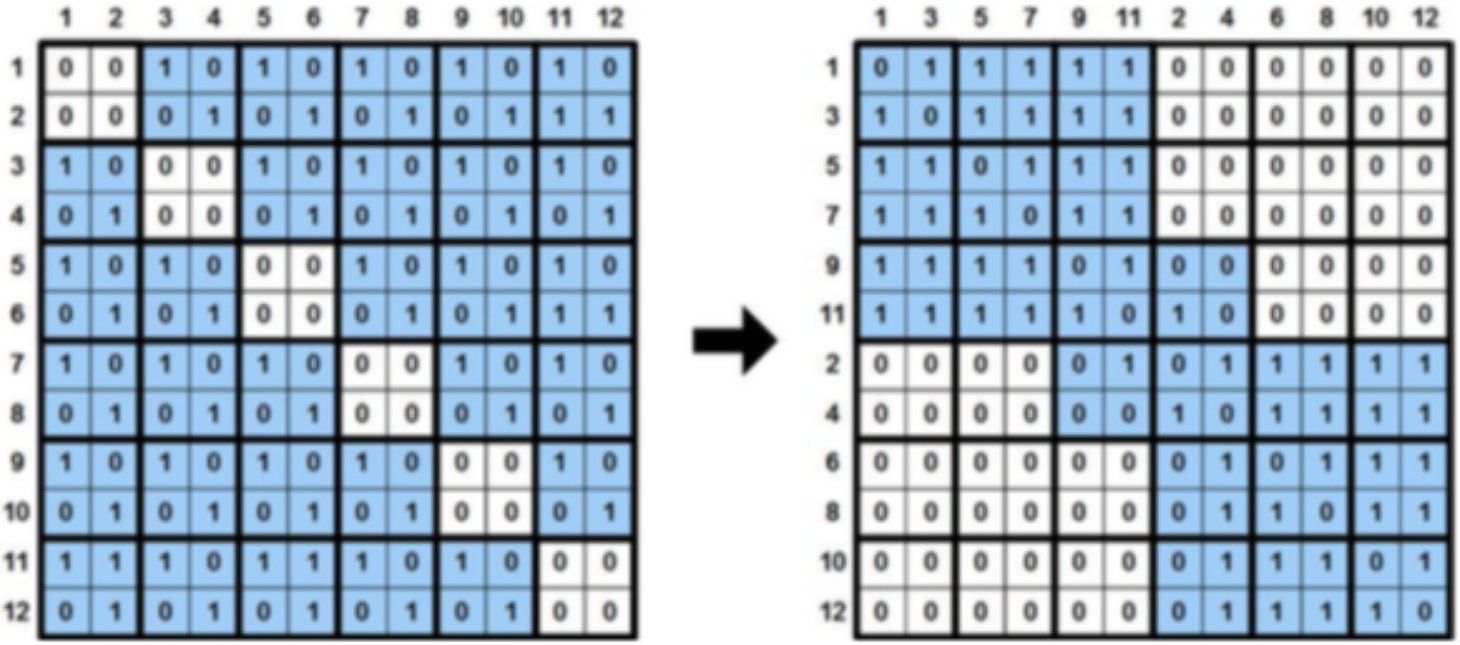


Fig 10 : Star

Fig 11 : Chain

Vocabulary based Graph Summarization



	1	2	3	4	5	6	7	8	9	10	11	12
1	0	0	1	0	1	0	1	0	1	0	1	0
2	0	0	0	1	0	1	0	1	0	1	1	1
3	1	0	0	0	1	0	1	0	1	0	1	0
4	0	1	0	0	0	1	0	1	0	1	0	1
5	1	0	1	0	0	0	1	0	1	0	1	0
6	0	1	0	1	0	0	0	1	0	1	1	1
7	1	0	1	0	1	0	0	0	1	0	1	0
8	0	1	0	1	0	1	0	0	0	1	0	1
9	1	0	1	0	1	0	1	0	0	0	1	0
10	0	1	0	1	0	1	0	1	0	0	0	1
11	1	1	1	0	1	1	1	0	1	0	0	0
12	0	1	0	1	0	1	0	1	0	1	0	0

	1	3	5	7	9	11	2	4	6	8	10	12
1	0	1	1	1	1	1	0	0	0	0	0	0
3	1	0	1	1	1	1	0	0	0	0	0	0
5	1	1	0	1	1	1	0	0	0	0	0	0
7	1	1	1	0	1	1	0	0	0	0	0	0
9	1	1	1	1	0	1	0	0	0	0	0	0
11	1	1	1	1	1	0	1	0	0	0	0	0
2	0	0	0	0	0	1	0	1	1	1	1	1
4	0	0	0	0	0	0	1	0	1	1	1	1
6	0	0	0	0	0	0	0	1	0	1	1	1
8	0	0	0	0	0	0	0	1	1	0	1	1
10	0	0	0	0	0	0	0	1	1	1	0	1
12	0	0	0	0	0	0	0	1	1	1	1	0

Fig 13 : Bit Compression

Vocabulary based **G**raph Summarization

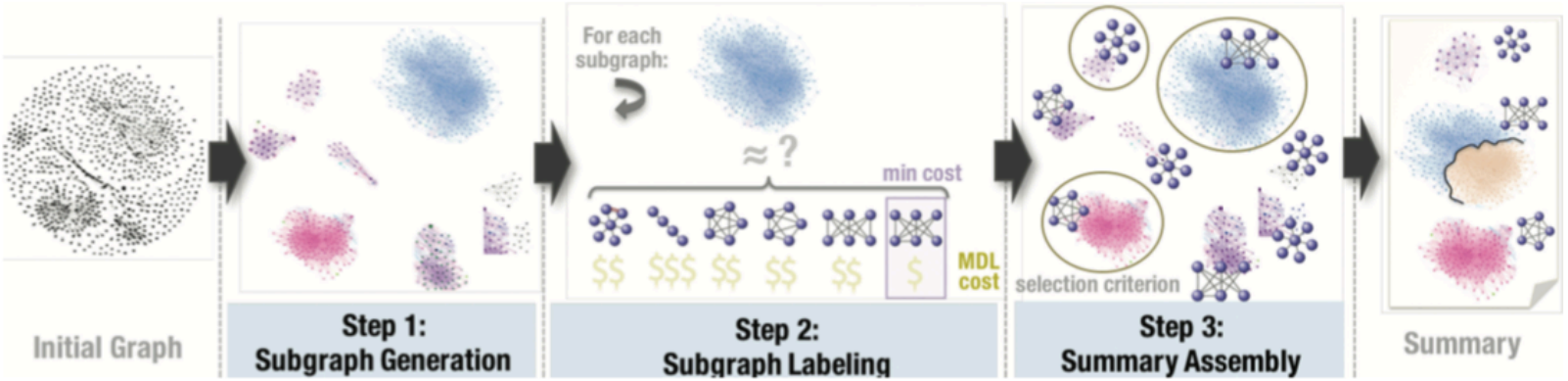


Fig 12. : **Vo**cabulary based Graph **S**ummarization

Graph Summarization

Modell

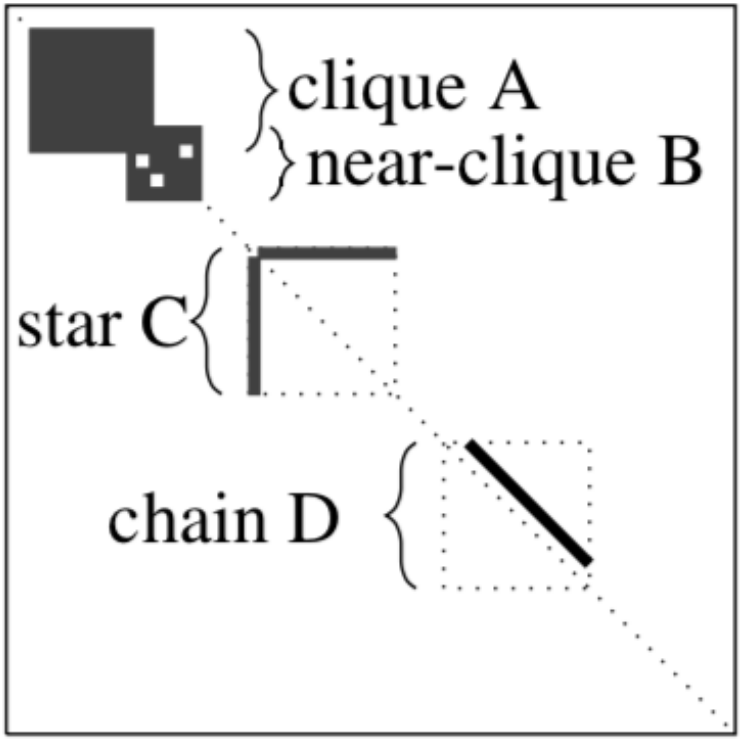


Fig 14 Ein Modell und Strukturtypen

Graph Summarization

Minimum Description Length : Modell

$$L(\mathrm{M}) = L_{\mathbb{N}}(|M|+1) + \log \binom{|M|+|\Omega|-1}{\Omega-1} +$$

Nummer der Strukturen

$$\Sigma_{s \in M}(-\log Pr(x(s)\,|\,M) + L(s))$$

Für jeden Strukturtypen in M

Graph Summarization

Minimum Description Length : Error

$$L(\mathsf{E}^+) = \log|\mathsf{E}^+| + ||\mathsf{E}^+|| * l_1 + ||\mathsf{E}^+||' * l_0$$

$$L(\mathsf{E}^-) = \log|\mathsf{E}^-| + ||\mathsf{E}^-|| * l_1 + ||\mathsf{E}^-||' * l_0$$

log(|E|) = die Kodierung der Anzahl von 1 in E

|E| = Die Anzahl aller 1 in E ,

|E|' = Die Anzahl aller 0 in E

l_1 = |E| / (|E| + |E|'),

l_0 = |E|' / (|E| + |E|')

Graph Summarization

VoG Problemdefinition

Gegeben sei ein Graph $G = (V, E)$,

Finde ein Modell M von G, mit dem

$$L(G, M) = L(M) + L(\mathsf{E})$$

den wenigsten Wert liefert

E = Adjazenzmatrix für Error = $\mathsf{M} \oplus \mathsf{A}$

Graph Summarization

VoG : Schritt 1 Teilgraphen generieren

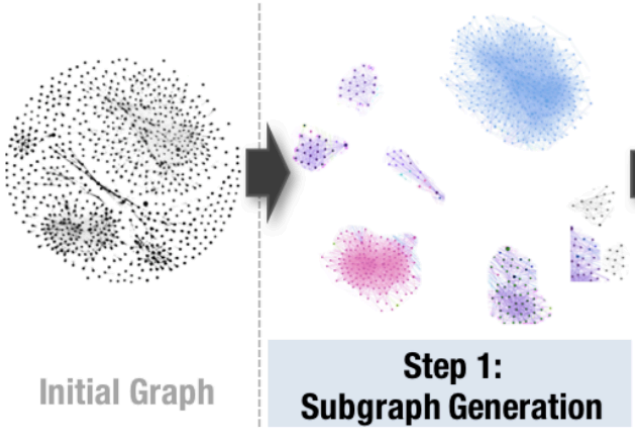


Fig 15 : Schritt 1 Teilgraphen generieren

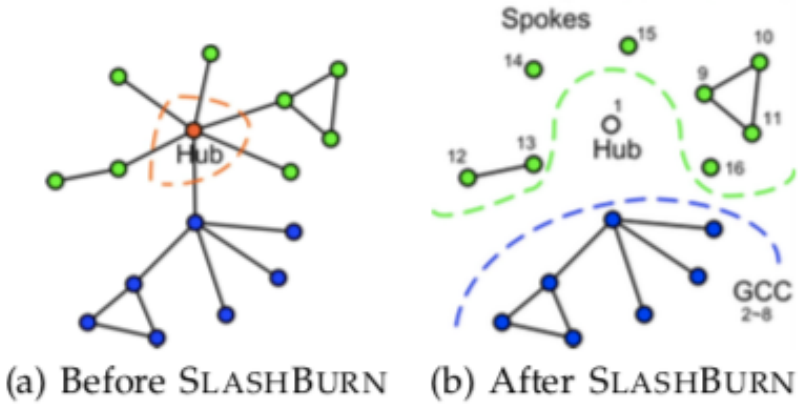


Fig 16 : SlashBurn

Graph Summarization

VoG : Schritt 2 Teilgraphen Labeling

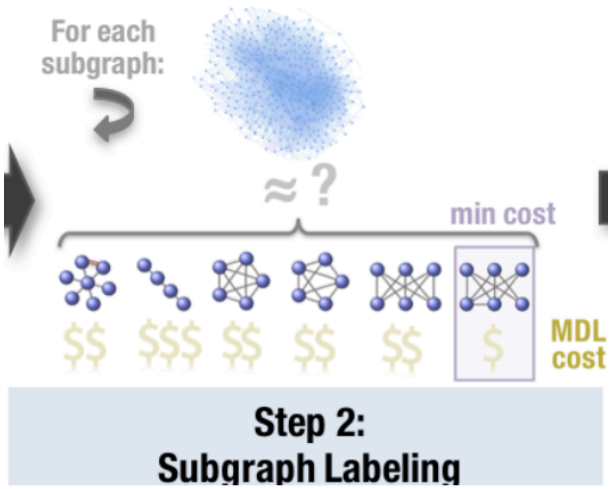


Fig 17 : Schritt 2 Teilgraphen Labellierung

$$L(G, m^*) = L(m^*) + L(E^+_{m^*}) + L(E^-_{m^*})$$

m^* = Teilgraph

$E^+_{m^*}$ = Überflüssige Kanten, die in m^*, aber nicht in G

$E^-_{m^*}$ = Fehlende Kanten, die in G, aber nicht in m*

Graph Summarization

VoG : Schritt 3 Summary Assembly

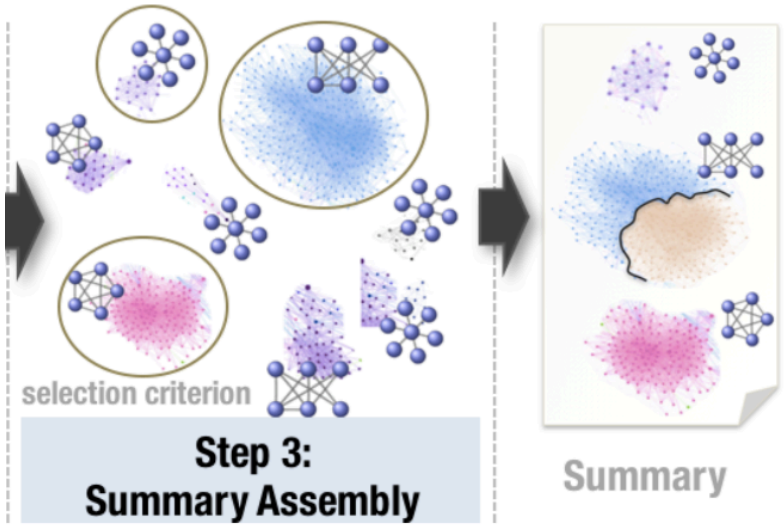


Fig 18 : Schritt 3 Summary Assembly

$$C = \{m^*_1, \ldots, m^*_n\}$$

PLAIN : $M = C$

TOP-K : $M = \Sigma_{i=1}^{k} m^*, k \leq n$

GREEDY'NFORGET

$= \forall s \in C$

$$L_{\mathbb{N}}(|M|+1) + \log \frac{|M|+|\Omega|-1}{\Omega-1} +$$

$$\Sigma_{s \in M}(-\log Pr(x(s)|M) + L(s))$$

Graph Summarization

VoG : Ergebnis (Beispiel)

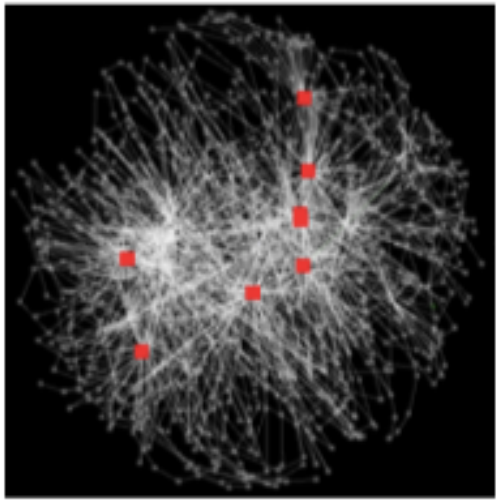


Fig 19 : Star in Wiki-Graphen

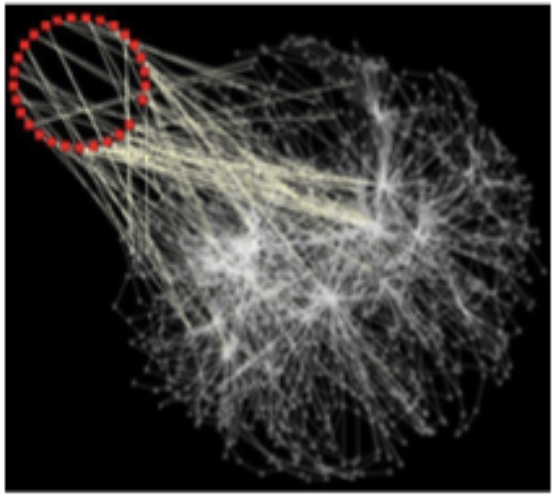


Fig 20 : Bipartite in Wiki-Graphen

Aktueller Zustand

Bis heute :

- Fertig dokumentierte Abschnitte
    - Einleitung, theoretische Grundlagen, Stand der Technik,

Plan

1. Exportieren der WossiDia Datensätze
2. Fertigung vom Dokumentieren :
    - Beschreibung des Algorithmus
    - Experiment
    - Ausblick
3. Eigene Idee : Implementierung von VoG mit Apache Spark

Diskussion und Q&A

Quelle

Fig 1, 14, 19, 20. VOG: Summarizing and Understanding Large Graphs (Danai Koutra, U Kang, Jilles Vreeken, Christos Faloutsos 2014)

Fig 2. : https://www.earthtouchnews.com/wtf/wtf/hairball-whodunit-guess-where-this-huge-hoghair-bezoar-came-from/

Fig 3 : Graph Summarization Methods and Applications: A Survey (Yike Liu, Tara Safavi, Abhilash Dighe, Danai Koutra 2016)

Fig 4,5 : GraSS : Graph Structure Summarization Kristen LeFevre, Evimaraia Terzi 2010

Fig 6,7,8,9,10,11 : StructMatrix: large-scale visualization of graphs by means of structure detection and dense matrices , Hugo Gualdron, Robson Cordeiro, Jose Rodrigues, June, 2015

Fig 12, 14, 17, 18 : VoG Summarizing and Understanding Large Graphs, Danai Koutra, U Kang, Jilles Vreeken, Christos Faloutsos, April, 2014

Fig 13, 16 : Slashburn, Youngsub Lim, U Kang, Christos Faloutsos, April, 2014