

Detekce komunit v sociálních sítích - algoritmus Nibble

Community detection in social media - algorithm Nibble

Daniel Loziňak

Semestrální práce

Vedoucí práce: RNDr. Eliška Ochodková, Ph.D.

Ostrava, 2022

Abstrakt

Táto semestrálna práca má úlohu opísať a implementovať lokálny zhukovací algoritmus Nibble v skoro lineárnom čase. Bol navrhnutý hlavne na detekciu komunit na sociálnych sieťach.

Klíčové slová

rez grafom, lokálny algoritmus, detekcia komunit

Abstract

The aim of this seminar paper is to describe and implement local clustering algorithm Nibble in near linear time. This algorithm was designed to detect communities on social media.

Keywords

graph cut, local algorithm, community detection

Obsah

1	Úvod	4
1.1	Zhlukovacie algoritmy	4
1.2	Algoritmus Nibble	4
2	Implementácia	5
2.1	Vodivosť	5
3	Záver	7

Kapitola 1

Úvod

1.1 Zhlukovacie algoritmy

Zhlukovacie algoritmy riešia úlohu rozdelenia objektov do skupín na základe podobnosti atribútov daných objektov. Platí pravidlo, že objekt musí patriť do skupiny a práve do jednej. Prístup k riešeniu problému závisí na konkrétnom algoritme. Rozoznávame dva hlavné typy zhukovacích algoritmov a práve aglomeratívne a divizívne. Pri aglomeratívnom zhukovaním vytvárame čoraz väčšie skupiny, až kým neexistuje len jedna skupina. Divizívne potom fungujú opačným spôsobom. Využitie zhukovacích algoritmov je mnoho. My sa budeme zaoberať detekciou komúní v sociálnych sieťach.

1.2 Algoritmus Nibble

S rýchlo rastúcim objemom sociálnych sietí ako sú Facebook, Instagram alebo Twitter je takmer nemožné vykonávať výpočty nad všetkými dátami, preto sa objavila otázka, a síce ako určiť zhuk ak budeme mať na vstupe práve jeden vrchol grafu. Tu využíva algoritmus Nibble metriku vodivosti, ktorá meria vnútorné a vonkajšie spojenia v zhuku. Ďalšia otázka je, v akom poradí budeme vrcholy prechádzať. Ako prvé nám môže napadnúť riadiť sa najkratšou cestou od vstupného vrcholu, no v sociálnych sieťach majú grafy nízky priemer, čo je pre prístup najkratšej cesty zlá voľba. Poradie vrcholov v algoritme Nibble závisí na krátkych náhodných cestách, ktoré vedú mimo zhuku. Krátka náhodná cesta je vlastne hrana vedúca mimo zhuku.

Grafy veľkých sociálnych sietí mávajú relatívne malý počet kvalitných zhukov, kde je najviac vhodný práve lokálny algoritmus. [1]

Kapitola 2

Implementácia

2.1 Vodivosť

Nech máme neorientovaný graf $G = (V, E)$. Zhluk G je podmnožina V taká, že má vysokú vnútornú, ale riedku vonkajšiu spojitosť so zvyškom grafu. Kvalitu zhľuku potom meriame vodivosťou, čo je podiel vonkajších spojení a celkovej veľkosti spojení. Vonkajšie spojenia si môžeme predstaviť ako všetky možné cesty zo zhľuku do zvyšku grafu. Vodivosť je teda pravdepodobnosť, že pri ceste jednou hranou grafu opustíme zhľuk.

Nech $d(i)$ je stupeň vrcholu i , potom pre $S \subseteq V$ definujeme objem S ako:

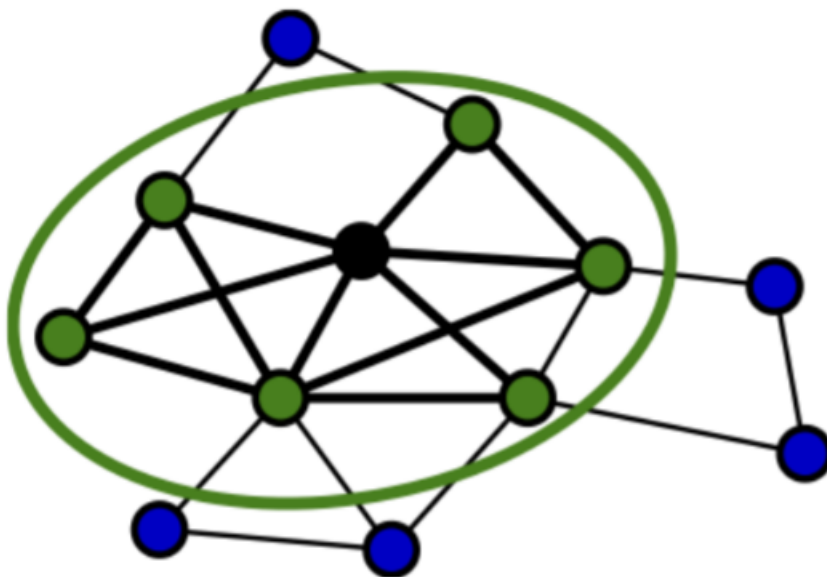
$$\mu = \sum_{i \in S} d(i)$$

Platí $\mu(V) = 2|E|$. Nech $E(S, V - S)$ je množina hrán spájajúcich vrcholy z S do $V - S$. Potom definujeme vodivosť ako $\phi(S)$ ako:

$$\phi(S) = \frac{|E(S, V-S)|}{\min(\mu(S), \mu(V-S))}$$

A vodivosť G ako:

$$\phi(G) = \min_{S \subseteq V} \phi(S)$$



Obr. 2.1: Zhluk

Ako vidíme na obrázku 2.1, počet možností opustenia zhľuku krátkym krokom je 7. Objem $\mu(S) = 33$ a Objem $\mu(V - S) = 11$, teda vodivosť $\phi(S) = 7/11$.

Kapitola 3

Záver

Bibliografia

1. DANIEL A. SPIELMAN, Shang-Hua Teng. *A Local Clustering Algorithm for Massive Graphs and its Application to Nearly-Linear Time Graph Partitioning*. Dostupné tiež z: <https://arxiv.org/pdf/0809.3232.pdf>.