Seminarski rad:

Rudarenje grafova u analizi socijalnih mreža

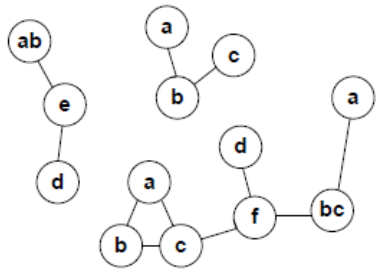
---

profesor: prof. dr Veljko Milutinović

student: Dušan Ristić 3020/2014

Uvod

Šabloni u grafovima su osnova za nekoliko ključnih aplikacija u grafovima, uključujući indeksiranje grafova, pretragu grafova, klasifikaciju grafova i klasterizaciju. Postojeći algoritmi za rudarenje grafova postigli su veliki uspeh koristeći strategije koje efikasno obilaze prostor šablona. Ipak, definicija čestih podgrafova možda nije odgovarajuća za nove slučajeve korišćenja koji se javljaju u socijalnim i informacionim mrežama. Kao prvo, definicija nije dovoljno elastična da bi uhvatila nejasne šablone koji se javljaju u masivnim atributivnim grafovima. Slika 1 pokazuje jedan primer gde je za svaki čvor povezan skup labela. Ove labele mogu biti filmovi preporučeni od strane nekog korisnika, funkcije koje prenosi gen ili upadi koje je inicirao kompjuter. Kao što se vidi na slici 1, a, b, c se često pojavljuju zajedno i formiraju asocijativni šablon, dok se c i d ne asociraju zajedno. Ipak, {a, b, c} nije ni česti podgraf niti čest skup elemenata ako svaki čvor tretiramo kao transakciju. Šablon {a, b, c} ima tri karakteristike: (1) Blizinu, ove 3 labele su usko povezane; (2) Učestalost, pojavljuju se mnogo puta; (3) Fleksibilnost, one nisu uvek povezane na isti način. Zbog ovih karakteristika, ne možemo primeniti tradicionalne algoritme za rudarenje čestih podgrafova kao što su FSG I gSPAN da bismo ih pronašli. Takođe, ne možemo koristiti ni rudarenje čestih skupova elemenata jer se {a, b, c} ne pojavljuju u istom skupu čvorova.



Slika 1 Šablon blizine {a, b, c}

Kao drugo, za male grafove kao što su hemijske strukture izomorfno proveravanje nikada ne predstavlja problem kao što pokazuju postojeći algoritmi za rudarenje čestih grafova. Ipak, za velike grafove kao što su upadne mreže i socijalne mreže može postojati ogroman skup izomorfnih uključivanja koja postoje za određeni česti podgraf. Postaje skupo da se generišu svi česti podgrafovi. Da bi se prevazišla navedena dva problema predložen je novi koncept šablona grafa koji se naziva šablon blizine. Šablon blizine je podskup labela koje se ponavljaju u više usko povezanih podgrafova u grafu. {a, b ,c} na slici 1 je primer. Šablon blizine je skup elemenata. Ipak, on ima jedan zahtev za povezivanje: Labele moraju biti asocirane usko i često u grafu. Na primer, u socijalnoj mreži, može biti skup filmova koji su pogledale više grupa korisnika. Dakle, da bi se pronašli šabloni blizine među filmovima, ne treba se uzeti u razmatranje samo skup filmova koje je pogledala svaka osoba ponaosob(u tom slučaju to bi bio tradicionalni problem rudarenja čestih elemenata), već treba razmotriti i filmove koje su pogledali prijatelji korisnika i prijatelji njegovih ili njenih prijatelja. U ovom slučaju labele asocirane za dva različita čvora su povezane zbog veze između ova dva čvora. Isti problem rudarenja je i pronalaženje povezanosti upada na internetu, gde svaki čvor predstavlja jednu IP adresu i postoji direktna veza između dve IP adresa ako se napad dešava između te dve adrese. Interesantno je naći povezanost različitih tipova napada, što se može koristiti u analizi upada.

U ovom radu, prvo uvodimo intuitivni model susedne asocijacije da bismo definisali i alocirali šablone blizine tako što ćemo identifikovati njihova uključenja u grafu, a potom naći težinski maksimalni nezavisni skup među ovim uključenjima. Iako je ovaj pristup intuitivan, on nije efikasan za nalaženje šablona u velikim grafovima zbog kompleksnosti nabrajanja uključenja i pronalaženja maksimalnog nezavisnog skupa. Zbog toga redefinišemo šablone blizine sa tačke gledišta uticaja, koristeći propagacioni model probabilističkih informacija. Na osnovu ovog modela, predlažemo nove tehnike za pronalaženje šablona blizine u velikim grafovima, koje razmatraju uslovnu probabilističku asocijaciju labela u svakom čvoru.

Modelujemo problem određivanja blizine među labelama u dva odvojena pristupa, susedna asocijacija i propagacija informacija. Iako je model susedne asocijacije direktan pristup pronalaženja povezanosti među labelama zasnovan na njihovoj udaljenosti preko veza grafa, pokazalo se da je ovaj pristup neefikasan za velike grafove. U modelu propagacije informacija, razvijene su nove probabilističke tehnike za određivanje blizine među labelama u grafu.

Osnovno

Graf sa atributima $G = (V, E)$ ima skup labela L i za svaki čvor je povezan skup labela. Skup labela čvora u u grafu G je $L(u)$. Neka je I podskup labela takav da su labele u I usko povezane i ponavljaju se u grafu G. Tada se I naziva šablonom blizine. Šabloni

blizine se mogu spustiti na nivo čestih skupova elemenata ako se izbace sve veze iz *G*.

Neka je $D = \{t_1, t_2, \ldots, t_m\}$ skup nezavisinih transakcija (u kontekstu grafa sa atributima, skup čvorova). Svaka transakcija sadrži podskup elemenata u *L*.

Definicija 1 (Podrška). *Podrška sup(I) skupa $I \subseteq L$ je broj transakcija u skupu podataka koji sadrži I. Ponekad se koristi i procenat da se izrazi podrška.*

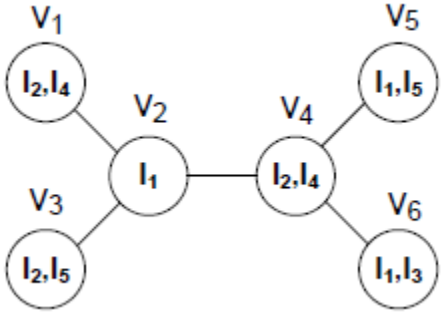
Skup elemenata se naziva čestim ako je njegova podrška veća od korisnički definisanog minimalnog praga. Skoro svi klasični algoritmi rudarenja čestih skupova elemenata imaju osobinu zatvaranja na dole kako bi smanjili prostor pretrage šablona.

Definicija 2 (Zatvaranje na dole). *Za čest skup elemenata, svi njegovi podskupovi su česti; odatle sledi da za bilo koji skup koji nije čest svaki njegov nadskup takođe nije čest.*

Nažalost, pošto rudarenje čestih skupova elemenata ne razmatra veze u atributivnom grafu, ono može da propusti interesantne šablone. Na slici 2 se može videti primer toga. Ako svaki čvor posmatramo kao nezavisnu transakciju, $\{l_1, l_2\}$ neće biti prepoznato kao čest podskup. Ova dva elemeta se ne pojavljuju zajedno ni u jednom od čvorova. Ipak, pažljivim posmatranjem slike 2 možemo zaključiti da se oni uvek javljaju na razdaljini dužine jedan. $\{l_1, l_2\}$ je šablon blizine: l_1 se asocira za blizinu l_2.

Za šablon blizine *I*, moramo da identifikujemo lokacije ovog šablona u *G*. Svaka od ovih lokacija mora da sadrži sve labele iz *I*.

Definicija 3 (Uključenje i mapiranje). *Dat je graf G i podskup njegovih čvorova π, $\pi \in V(G)$. Neka je $L(\pi)$ skup labela u π, $L(\pi) = U_{u \in \pi} L(u)$. Ako je dat podskup labela I, π se naziva uključenjem od I ako je $I \subseteq L(\pi)$. Mapiranje f između I i čvorova u π je funkcija $\varphi: I \rightarrow \pi$, $\exists l$, $\varphi(l) \in \pi$ and $l \in L(\varphi(l))$. Mapiranje je minimalno ako je surjektivno, $\forall\ v \in \pi$, $\exists\ l$ odnosno $\varphi(l) = v$.*



Slika 2 Čest skup vs Šablon blizine

Na slici 2, $\{v_1, v_2, v_3\}$ formira uključenje $\{l_1, l_2, l_5\}$. Postoje dva moguća mapiranja u ovom uključenju: (1) φ_1 mapira l_1 na v_2, l_2 na v_1 i l_5 na v_3, i (2) φ_2 mapira l_1 na v_2, l_2 na v_3 i l_5 na v_3. Kod ova dva mapiranja φ_1 je minimalno, a φ_2 nije. Čvorovi u π mogu biti nepovezani. Na primer, $\{v_1, v_3\}$ je uključenje $\{l_4, l_5\}$.

Ako je dat skup elemenata *I* I mapiranje φ, potrebna nam je funkcija $f(\varphi)$ koja će meriti njegovu jačinu asocijativnosti: koliko jako su povezane mapirane labele u π. Na primer $f(\varphi)$ može da bude inverzni dijametar od φ. Pošto može da postoji više mapiranja u π, uvek biramo ono sa najvecom vrednošću $f(\varphi)$. Da bismo pojednostavili zapis, dalje ćemo koristiti $f(\pi)$ kao jačinu uključenja.

U sledećem delu ćemo ispitati 2 modela da bismo definisali podršku šablonima blizine.

Model susedne asocijacije

Složenost šablona blizine povećava se interkonekcijama labela u grafu. Treba uraditi sledeća 3 koraka da identifikuje šablone blizine:

Korak 1. Naći sva uključenja, $\pi_1, \pi_2, \ldots, \pi_m$ skupa elemenata *I* u grafu,

Korak 2. Za svako uključenje π, izračunati njegovu jačinu $f(\pi)$,

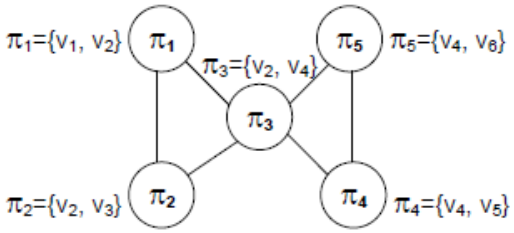
Korak 3. Agregirati jačinu uključenja, $F(I) = \sum_{i=1}^{m} f(\pi_i)$. Uzeti $F(I)$ kao podršku *I*.

Da bismo našli podršku šablona blizine, treba prvo pobrojati sva uključenja tog šablona. Na žalost, zbog veza u grafu, može biti eksponencionalan broj redundantnih uključenja. Kao prvo, granica između uključenja šablona nije očigledna. Kada se dva uključenja preklapaju, deo koji se preklapa će se brojati duplo. Podrška izračunata iz višestrukih uključenja bi narušila osobinu zatvaranja na dole

(Definicija 2). Podrška šablona I može biti manja nego podrška šablona I', iako je $I \subseteq I'$, zbog čega je teško dizajnirati brze algoritme za rudarenje. Kao drugo, bilo koji podskup čvorova, π, može biti uključenje šablona I sve dok $I \subset L(\pi)$, ali za slabo povezana uključenja, njihova jačina može biti zanemarljiva.

Da bismo rešili ova dva problema, uvodimo dva modela, model susedne asocijacije i model propagacije informacija.

Neka su $\pi_1, \pi_2, \ldots, \pi_n$ uključenja I u G. Pravimo preklapajući graf: svaki čvor predstavlja uključenje, a grana povezuje dva uključenja ako oni sadrže bar jedan isti čvor. U preklapajućem grafu svaki čvor ima težinu $f(\pi)$. Slika 3 pokazuje primer grafa delimičnog preklapanja izvedenog iz slike 2.



Slika 3 Preklapajući graf

Za rudarenje čestih grafova u nekom grafu predloženo je da se koristi maksimalni nezavisni skup kao podrška podgrafa, za šta je pokazano da ima osobinu zatvaranja na dole. Nezavisan skup u grafu je podskup čvorova koji nemaju vezu između sebe. Na slici 3 uključenja π_1, π_4 formiraju maksimalni nezavisan skup. Ovaj concept se može proširiti na preklapajući težinski graf. Za skup labela I, podrška I bi mogla biti suma težina čvorova podeljena sa maksimalnom težinom nezavisnog skupa. Ovaj model se naziva modelom susedne asocijacije.

Iako model susedne asocijacije rešava problem preklapanja šablona, on je generalno NP-težak uzimajući u obzir broj uključenja za dati šablon. Pošto taj broj može biti veliki, nije praktično da se generišu sva uključenja šablona blizine I onda nađe njihova maksimalna težina nezavisnog skupa. Zato pristupamo drugom modelu, modelu propagacije informacija.

Model propagacije informacija

Model susedne asocijacije ispituje asocijativnost iz perspective strukture grafa. Na primer, za dve labele l_1, l_2, u grafu, koliko čvrsto su povezane i koliko često su povezane. Moguće je ispitati isti problem iz perspektive uticaja mreže. Uzmimo kao primer socijalnu mrežu za preporučivanje filmova, gde korisnici mogu da preporučuju filmove svojim prijateljima. Pretpostavimo da je G_0 inicijalni graf. Na osnovu preporuka, korisnici mogu pogledati još filmova i generisati novi graf G_1 sa ažuriranom listom pogledanih filmova. Ovaj process iterira sve dok ne dostigne stabilni graf gde se lista filmova za bilo kog korisnika više ne menja.

$$G_0 \rightarrow G_1 \rightarrow \ldots \rightarrow G_n$$

U idealnoj situaciji, bilo bi značajno da rudarimo česte skupove elemenata u G_n. Ipak, u stvarnosti imamo samo nekompletnu sliku grafa između G_0 i G_n. Šabloni blizine u G_i mogu se tumačiti kao aproksimacija čestih skupova podataka u G_n. Ako možemo da simuliramo uticaj tako što generišemo $\tilde{G}$ iz G_i da aproksimira G_n, možemo onda koristiti česte skupove podataka izrudarene u $\tilde{G}$ da predstavljaju šablone blizine iz G_i. Ovo je osnovna ideja modela propagacije informacija.

Dati graf se smatra za trenutno stanje, i asocijacije među labelama u budućem stanju se dostižu iterativnim stohastičkim procesom. Neka je $L(u)$ trenutno stanje u, opisano labelama prisutnim u u, i neka je l distinktna labela propagirana od strane jednog od suseda i $l \notin L(u)$. Odatle, verovatnoća da posmatramo $L(u)$ i l je

$$P(L \cup \{l\}) = P(L|l)P(l)$$

gde je $P(l)$ verovatnoća da se l nalazi u susedima od u, a $P(L|l)$ verovatnoća da je l uspešno propagiralo do u.

Za više labela, $l_1, l_2, \ldots, l_m$, zajednička verovatnoća za posmatranje

$$P(L \cup \{l_1, \ldots, l_m\}) = P(L|l_1)*\ldots*P(L|l_m)*P(l_1)*\ldots P(l_m).$$

Propagacioni model hvata bitnu karakteristiku u društvenim grafovima gde čvorovi mogu uticati jedni na druge. Što se razdaljina povećava, to je uticaj manji, a to je baš ono što šabloni blizine žele da uhvate. U sledeća dva odeljka su prikazana dva odvojena pristupa za dodelu vrednosti prethodno pomenutim uslovnim verovatnoćama, $P(L|l)$, uz detaljne algoritme. Ova dva pristupa obrađuju

situaciju gde se ista labela propagira iz više čvorova sa različitim razdaljinama.

Najbliža probabilistička asocijacija

U modelu najbliže probabilističke asocijacije (NPA), uslovna verovatnoća $P(L(u)|l)$ u daljem tekstu $A_u(l)$ se definiše kao:

Definicija 4 (Najbliža asocijacija) *Neka je l labela prisutna u v koji je najbliži u, gde $l \notin L(u)$. $A_u(l) = e^{-\alpha d}$, gde je d udaljenost od v do u, a α je konstanta raspada ($\alpha > 0$).*

$A_u(l)$ raspada se na nulu kako se d približava ∞. Za bežtežinski graf uzimamo da je $d = 1$ za svaku vezu. Algoritam za pronalaženje stabilnog propagiranog grafa $\tilde{G}$ je sledeći:

Ulaz: Graf G, parametar odsecanja ε.
Izlaz: Međuskup podataka $\tilde{G}$.

1: $i = 0$
2: for all čvorovi u u G do
3: Neka je $L_0(u)$ skup labela čvora u
4: $\forall\ l \in L_0(u)$, $A_u(l) = 1$; inače $A_u(l) = 0$
5: end for
6: for all čvorovi u u G do
7: for all labele l u $L_i(v) \setminus L_i(u)$, v je sused od u do
8: $A_u(l) = e^{-\alpha d}$ (izabrati maksimalnu vrednost)
9: Ako je manje od ε ne propagirati l do u
10: end for
11: $L_{i+1}(u) = \{L_i(u) \cup \{l\} | A_u(l) > 0\}$
12: end for
13: if $L_{i+1} = L_i$ za sve čvorove u G then
14: Izdaj A_u za sve $u \in V(G)$
15: else
16: $i = i + 1$, vratiti se na korak 2
17: end if

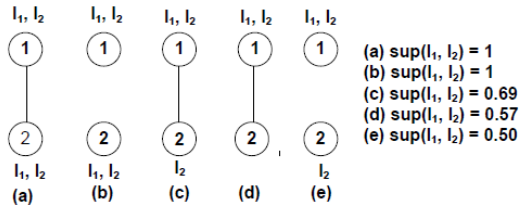
Vrednost asocijativnosti bi trebalo da se smanjuje što se rastojanje među čvorovima povećava, pa zbog toga postoji parametar odsecanja ε. Ne propagiramo labelu kada je najbliža vrednost asocijacije manja od ε.

Primetite da je $A_u(l) = 1$ kada sam čvor u sadrži labelu l; $A_u(l) = 0$, kada je čvor u značajno udaljen od čvora sa labelom l ili ne postoji putanja između u i nekog čvora koji ima labelu l. Pošto se asocijacija određuje po najbližem pojavljivanju labele, ovaj model nazivamo najbliža asocijacija. Kada se napravi međuskup podataka $\tilde{G}$ možemo definisati podršku šablona blizine.

Definicija 5 (Probabilistička podrška). *Ako je dat međuskup podataka $\tilde{G}$ izveden modelom najbliže probabilističke asocijacije, podrška $I = \{l_1, l_2, \ldots, l_m\}$, $sup(I) = \frac{1}{V}\sum_{u \in V} A_u(l_1) \ldots A_u(l_m)$ gde je $A_u(l)$ verovatnoća posmatranja l u u.*

Definicija podrške u NPA ima osobinu zatvaranja na dole, $sup(I) \geq sup(J)$ ako je $I \subseteq J$.

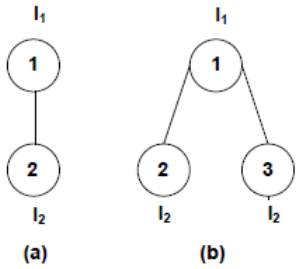
Definicija je takođe u skladu sa definicijom podrške čestih skupova elemenata, gde $A_u(l)$ može biti samo 1 ili 0. Na slici 4 se vidi primer kada je konstanta raspada jednaka 1.



Slika 4 Konzistencija: NPA i čestih skupova elemenata

Opadajuća vrednost podrške oslikava snagu asocijacije između l_1 i l_2 u različitim strukturama.

NPA model je brz za računanje, ali postoji potencijalni problem. Za bilo koji čvor razmatra se samo najbliži sused za svaku labelu, pa ne može da razdvoji situacije kada postoje više od jednog najbližeg čvora sa istom labelom. Slika 5 prikazuje taj primer.



Slika 5 Problem sa NPA

Slika 5 prikazuje 2 grafa i u oba slučaja je $sup(l_1, l_2) = 0.37$ po NPA. Da bismo razdvojili ova dva slučaja moramo uvesti poboljšanje.

Normalizovana probabilistička asocijacija (poboljšanje)

U normalizovanom probabilističkom asocijativnom modelu (NmPA), uzimamo u obzir sva najbliža pojavljivanja iste labele.

Definicija 6 (Normalizovana asocijacija). *Ako je dat atributivni bestežinski graf G i čvor u i ako je broj suseda čvora u jednak n, i ako ima m suseda sa labelom l, normalizovana probabilistička asocijacija l u u je* $NA_u(l) = P(L(u)|l) = \frac{m}{n+1}e^{-\alpha}$.

Normalizacioni faktor $Z = \frac{m}{n+1}$ će dati veću asocijativnu snagu labelama koje sadrže mnogi susedi. Da bismo razlikovali 2 slučaja sa slike 5 biramo *n* + 1 kao delilac umesto *n*.

Primenjuje se isti algoritam kao i kod NPA, sa tim da se u liniji 8 sada ubacuje nova jednačina za ažuriranje verovatnoće:

$$NA_u(l) = \frac{1}{n+1}\sum_{v \in \mathrm{N(u)}} e^{-\alpha} * NA_u(l)$$

gde je $NA_u(l)$ je asocijativna snaga l kod v, a N(u) je skup suseda čvora u. Pošto l može da bude labela koja je propagirala iz nekog drugog čvora $NA_u(l)$ može biti manje od 1.

Zaključak

Uvedeni su novi koncepti za šablone u grafovima – šabloni blizine, koji predstavljaju poprilično odvajanje od tradicionalnih koncepata čestih podgrafova i čestih skupova elemenata. Šablon blizine ublažava granicu između skupa elemenata i strukture. Diskutovano je o slabostima modela susedne asocijacije i predloženo je bolje rešenje u vidu modela propagacije informacija koji može da transformiše složen problem rudarenja u uprošćeni problem rudarenja težinskih skupova elemenata, koji se dalje jednostavno rešava korišćenjem izmenjenog FP-tree algoritma. Opisana je funkcija koja predstavlja meru interesantnosti pronađenih šablona. Sve u svemu, predloženi su svi koraci za rudarenje novih šablona blizine.

Reference

- A. Khan, X. Yan, and K.-L.Wu. Towards proximity pattern mining in large graphs. In *SIGMOD, 2010.*
- C.C. Aggarwal, Y. Li, J. Wang, J. Wang. Frequent Pattern Mining with Uncertain Data. KDD 2009.
- Presentation on Frequent Pattern Growth (FP-Growth) Algorithm: An Introduction by Florian Verhein
- J. Han, J. Pei, Y. Yin. Mining frequent patterns without candidate generation. SIGMOD, 2000.