Analiza linkova

Analiza socijanih mreza I nacini obradjivanja I rangiranja

Ranije se koristilo - poredjenje sadrzaja upita sa sadrzajem indeksiranih stranica, metode imaju koren u data majningu I zasnovane na bulovom I vektorskom prostoru, tfidf. Rast podataka je danas veci pa se ove metode manje koriste. Ove metode nisu lose, imaju znacajnu pirmenu ali u ovakvom okruzenju nisu najpogodnije. Npr spamovali stranice da se najbolje rangiraju.

Osnova pretrage I analize weba – analiza direktivnosti hiperlinkova, moguce je raydvojiti znacajne od manje znacajnih stranica, uoliko veliki broj stranica ukazuje na jednu stranicu njen autoritet raste I takvo ponassanje dovelo je do toga da se analiziraju hiperlinkovi, odnosno za konkretan upit imaju najvise smisla. Page rank I Hits se zasnivaju na ovim metodama.

Linkovi se mogu iskoristiti za klasterizaciju ljudi I sajtova na vebu, moguce je ovojiti zajednice ljudi koje su povezane na webu I formiranje konteksta sto void kategorizaciji veba.

# Analiza socijalnih mreza

Analiza zajednica, najveci znacaj je identifikacija korisnika I njihovih veza. Web svet za sebe, izgleda kao jedan graf gde su ljudi I organizacije medjusobno povezani. Svaka veza je hyperlink izmedju dva entiteta. Njihovim analiziranjem se dobija uloga, pozicija ili znacaj, sto je najbitnije, pojedinca ili organizacije.

Dva nacina analiziranje mreze na vebu – posmatra se centralnost I prestiz nekog entiteta (pojedincana str, sajt ili grupa).

**Centralnost**

Neko ko ima visok stepen centralnsoti znaci da ima veliki broj veza, na taj nazin je znacajniji. Stepen centralnosti Cp(i)=d(i)/n-1 kod neusmerenog grafa, kod usmerenog gledaju se samo odlazne veze kod nekog cvora.

Centralnost moze da bude posredna, udaljenost pojedinacnog cvora od ostalih svorova u mrezi (closeness centrality). Kod neusmerenog grafa stepen je vr koja kaze da vr jedan moze da se dobije kada je cvor vezan sa svim ostalim cvorovima u mrezi, u suprotnom n-1 podeljeno sa \*.

Betweenness centrality

Posrednost ima za cilj da identifikuje Koliko veza prolazi preko pojedinacnog entiteta, znacaj je veca ako se prolazi preko tog cvora ako treba da se ostvari veza preko nekih cvorova. Moze postojati veci broj najkracih puteva za dva cvora, neki prolaze kroz cvorove za koje racunamo centralnost, a neki ne.

**Prestiz**

Zamisljen kao mreza koja treba da omoguci precizniju analizu mreze. Razlika u odnosu na centralnost jeste da se kod proracuna racunaju samo dolazni linkovi.

Koriste se tri mere:

Stepen prestiza je zasnovan na tome da se posmatra koliko ostalih cvorova direkto ukazuje na cvor koji analiziramo. Pn = d(i)/n-1, n broj cvorova u grafu.

Proximity prestiz se definise kao rastojanje pojedinacnog cvora u odnosu na sve ostale entitete koji kofigurisu u mrezi. Hocemo da odredimo koliko je svaki cvor u mrezi daleko od cvora koji trenutno posmatramo. Suma rastojanja podeljenja sa velicinom seta dobija se znacaj, sto vise cvorova ukazuje na taj cvor njegov znacaj je veci.

Rank prestiz iskoriscen za Page rank alg. Zasnovan na pojmu znacaja pojedinacnih ucensika u mrezi. Uveden je concept glasanja, posmatraju se samo dolazni linkovi. Za svaki cvor odredi se njegov znacaj sto predstavlja sumu znacaja svih cvorova koji ukazuju na taj cvor. Tezina linka odredjena je prestizom cvora. Matrica susednosti se koristi. Bitno je da se prvi put uvodi konstatacija da se koristi znacaj pojedinacnog cvora za racunanje znacaja trenutnog znacaja koji na njega pokazuju.

# Kocitiranost I bibliografsko uparivanje (bitno – moze tema da bude https://www.elsevier.com/solutions/sciencedirect/support/api)

Koriste se u analizi radova. Pronalazenje relevantnih radova. Dve mere se koriste, svaka u Hitsu iskoriscene.

**Ko-cititranost**

Oba rada citirana od rada k to znaca da ta dva rada mogu da budu povezana koju nije moguce direktno saznati, to je posredna relacija. Sto vise radova citita te radove njohova veza je znacajnija. Ako se napravi matrica cititranosti L, ima vr 1 ukoliko rad I citira j, a ako ne cititra 0.

**Bibliografsko uparivanje**

Uparivanje radova koji citiraju iste radove.

**PageRank i HITS**

Modelovanje nacina analize linkova na webu. Objavljeni 1998. PageRank doziveo vecu popularnost.

PageRank

Ne zavisi od upita. Rang prestiza je osnovna ideja oko koje je izgradjen PR. Koristi karakteristike weba u svojoj impl. Str je bitnija od ostalih str ukoliko bitnije str glasaju za nju. Glasa znaci da postoji link. Suma tezina str koje ukazuju na datu str je rank prestiza za nju.

Web posmatramo kao usmereni graf. Vrednost ranga odredjujemo tako sto sumiramo rangove str koje ukazuju na nju I to se podeli sa brojem str koje ukazuju na nju (valjda).

Matricna reprezentacija

Niz od n linearnih jna. Kompletan PG cele mreze moze da se predstavi kao vector. P = A transponovano \* P. Mat A je susednost cvorova u grafu.

L\*P=A\*P, ako je L=1 mozemo da je predstavimo na ovaj nacin I moze da se iskoristi iterativna metoda izracunavanja.

Web posmatramo kao Markovljev lanac, zasnovan na tome da svaka stranica se predstavlja kao cvor odnosno stanje. Odlazak sa jedne na bilo koju drugu str slucajna stvar. Onda je verovatnoca tranzicije 1/br. odlaznih linkova.   
Stohasticki process predstavljen stranicama (valjda). Na svakoj str imamo def vr prelaska na drugu str. Za svaku str mora da bude def vr da bismo mogli da imamo lanac. Iz bilo kog stanja prelazimo u bilo koje stanje ne zavisi od toga kako smo u to stanje dosli. Kada sumiramo vr po kolonama u matrici trebalo bi da se dobije jedinica.

Matrica A mora da bude nesvodljiva I aperiodicna. Ako ne zadovoljava ova dva svojstva ne mozemo da kazemo da L 1 pa ne moze PG da se racuna iteracijama. A da bude stohasticka. Ona na mnogim mestima ima nule. Ideja je da se matrica dopuni verovatnocama. Ako ne postoji direktna veza uvodi se teleport.

Markovljev lanac je aperiodicna ukoliko je k=1 za svako stanje.

Kako odrediti parameter za teleport? Ima formula sa dumping faktorom.

Prednosti: moguce izracunati vr ranga stranice u offline rezimu I ta vr nije uslovljena strukturom upita.

Kritika: ne moze da razdvoji autoritativnije str I str koje vise odg uslovima pretrage.

HITS (Hypertext)

Razlika u odnosu na PG je da je on query dependent. Radi prosirenje liste str koje mozda odg upitu. Kombinuje dve vr, authority I hub. Autoritet znaci da mnogo str ukazuje na neku str (ima puno dolaznih linkova). Hub je str koja ima mnogo odlaznih linkova, sa nje moguce doci na mnogo znacajnih str koje one opisuje.

Dobar hub ukazuje na mnogo dobrih autoriteta. I obrnutno, ukoliko na neku str ukazuje mnogo dobrih hubova ona je dobro autorizovana.

Upit se prosledi search engine I on odg nekim br str (npr 10). Taj skup str se prosiruje svim str na koje ukazuje bilo koja str iz osnovnog skupa I skupom str koje ukazuje na njih. Svakoj str se dodeli authority I hub score.

Dobro: moze da prosiri skup rezultata.

Problem: lako se spamuje tako sto se dodaju izlazni linkovi.