UNIVERSITY POLITEHNICA OF BUCHAREST
FACULTY OF AUTOMATIC CONTROL AND COMPUTERS
COMPUTER SCIENCE DEPARTMENT



# Data mining & Data warehousing

-Friend suggestion on social networks-

# Web Mining: SALSA

Enache Andrei-George

MTI1A

BUCHAREST

2022

# Cuprins

## 1. Importanta si aplicatiile practice ale algoritmului

Internetul reprezintă o colecție de informații nestructurate sau semi-structurate, în continua creștere, ce conține diverse surse și legături intre pagini. În ziua de azi, atunci când o persoana caută o informație pe internet, aceasta face o interogare în cadrul unui browser, fiindu-I returnate o serie de link-uri și pagini cu conținut asemănător, cel mai probabil, celui căutat. În cazul interogărilor pe subiecte generale pot rezulta multe pagini irelevante pentru utilizator, îngreunând astfel procesul de găsire a informațiilor utile [2].

Câteva dintre problemele întâlnite în cadrul căutărilor pe motoarele de căutare web sunt reprezentate de [7]:

- Interogarile restranse - informatia este foarte putina si greu de gasit, neputandu-se stabili concret daca este relevanta sau nu pentru utilizator;
- Interogari pe subiecte mari - exista foarte multe pagini ce dezbat un anumit subiect, neputandu-se stabili cu usurinta relevanta informatiilor afisate, utilizatorii avand nevoie doar de cateva site-uri pentru a-si satisface nevoie cautarilor;
- Sinonimia – cautarea anumitor informatii poate duce la gasirea unor pagini web cu tema similara, dar diferita de cee ace si-a dorit utilizatorul;
- Polisemia – rezultate din domenii diferite ale aceluiasi termen, putand fi catalogat si ca o cautare ambigua;
- Stiluri diferite ale autorilor – oameni din zone diferite ce folosesc un vocabular diferit pentru elemente din acelasi domeniu, putand induce in eroare utilizatorul.

Din fericire, una dintre cele mai importante caracteristici ale site-urilor web o reprezintă existenta legăturilor către alte site-uri, oferind astfel o structura generala asemănătoare intre diferite site-uri, putând astfel face generalizări cu scopul de a îmbunătăți calitatea elementelor afișate în urma căutărilor în browser, utilizând diverse metode de web mining.

Web mining-ul reprezintă procesul de aplicare a algoritmilor și tehnicilor de căutare a informațiilor pentru a extrage datele direct din paginile web. Printre principalele categorii de web mining se Numără:

- Web content mining – procesul de gasire a informatiilor utile din continutul text si media a paginilor web;

- Web structure mining – procesul de analizare a structurii paginilor web, a nodurilor de legatura si a conexiunilor intre diferite pagini;

- Web usage mining – procesul de extragere a informatiilor despre modul cum utilizatorii folosesc un anumite site.

În cadrul acestei lucrări a fost implementat algoritmul de web structure mining SALSA (Stochastic Approach for Link-Structure Analysis) cu scopul de a face o analiza a importantei anumitor pagini web în funcție de structurile hyperlink-urilor existente. Acest algoritm aeste utilizat pentru a extrage modele din hyperlink-urile web și se focuseaza pe acelea care unesc pagini diferite (Inter-Document Hyperlink), nu și acelea care trimit utilizatorul către aceeași pagina (Intra-Document Dyperlink). Scopul implementării acestui algoritm este de a spori relevanta rezultatelor căutărilor în motoarele de căutare, dar poate fi utilizat și în alte scopuri, cum ar fi sugerarea elementelor ce au legătură cu utilizatorul (ex: sugestia prietenilor pe facebook).

## 2. Prezentarea generală a algoritmului

SALSA (Stochastic Approach for Link-Structure Analysis) reprezintă un algoritm de clasificare a paginilor web ce le atribuie acestora scoruri de butuci (hubs) și de autorități (authorities), pe baza hyperlink-urilor și legăturilor dintre acestea [1].

Aceasta solutie a fost inspirate din alti 2 algoritmi de web structure mining, HITS si PageRank. La fel ca si HITS, SALSA confera paginilor web scoruri in functie de relevanta acestora in functie de un anumit subiect. Scorurile pe care le poate primi un site sunt denumite Buturugi (Hubs) si Autoritati

(Authorities). O autoritate mai mare reprezinta o relevanta mai mare a unui site asupra unui anumit subiect, avand mai multe hyperlink-uri ce tintesc acea pagina. Cu cat valoarea hub-ului unei pagini este mai mare, relevanta acelei pagini scade, fiind considerate doar o pagina ce trimite utilizatorul catre un site mai important. Asemanarea dintre acest algoritm si PageRank este modalitatea de calculare a scorurilor hubs si authorities prin intermediul unui drum la intamplare printr-un lant Markov ce reprezinta graful paginilor web.

În prima etapa a algoritmului se allege un subiect care este căutat prin intermediul unui motor de căutare. Din rezultatele găsite se vă alege un număr de pagini care vor fi verificate ulterior, luandu-se de asemenea hyperlink-urile prezente pe paginile alese, dar și cautandu-se hyperlink-uri ce duc către acele site-uri.

Algoritmul pornește de la ipoteza că paginile cu nivel de autoritate mai mare asupra unui subiect “t” vor fi vizibile de pe mai multe pagini din graful “C” (format din noduri ce reprezintă site-urile, marginile fiind considerate că fiind link-urile dintre pagini). Utilizând graful direcționat format din paginile rădăcină C se formează un graf bipartit nedirectionat “G”, având în componenta să subgraful “H” ce vă conține nodurile hub (care conțin hyperlink-uri ce redirectioneaza utilizatorul către alte pagini) și subgraful “A” ce vă conține nodurile autoritate (pagini către care utilizatorul este trimis prin apăsarea link-urilor de pe paginile hub)(Fig.1, Fig.2)[1].

$$—V_h = \{s_h \mid s \in \mathcal{C} \text{ and } out\text{-}degree(s) > 0\} \text{ (the } hub \text{ side of } \tilde{G}).$$

$$—V_a = \{s_a \mid s \in \mathcal{C} \text{ and } in\text{-}degree(s) > 0\} \text{ (the } authority \text{ side of } \tilde{G}).$$

$$—E = \{(s_h, r_a) \mid s \to r \text{ in } \mathcal{C}\}.$$

*Fig.1. Împărțirea nodurilor inițiale în subgrafurile H și A*

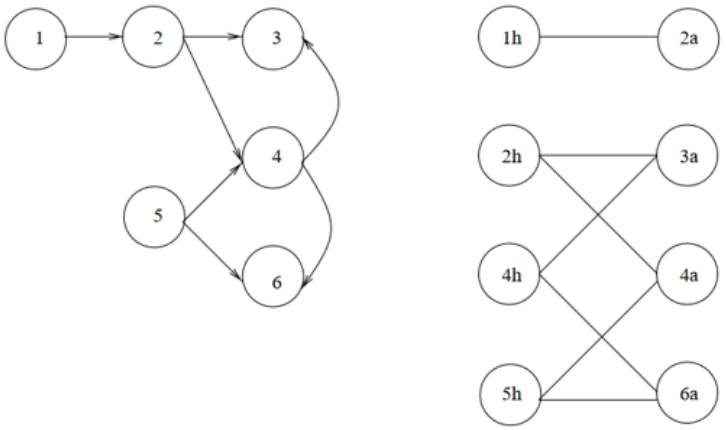


*Fig.2. Transformarea grafului directionat “C” in graficul bipartit “G”[1]*

Unde:

- G = Graficul bipartite
- S = pagini web neizolate apartinand lui “C”, reprezentate de 2 noduri ale lui “G” (“sh” si “sa”)
- Vh = marginile care ies din hub si redirectioneaza utilizatorul catre o autoritate
- Va = marginile care tintesc spre autoritate
- E = marginile nedirectionate

**a) Metoda de implementare 1**

După organizarea subgrafurilor este necesara realizarea a doua drumuri aleatorii distincte, fiecare dintre acestea pornind dintr-o latura a grafului, luandu-se astfel în considerare un lanț de deplasare al hub-urilor și unul al autorităților. Pentru a realiza stările de tranziție ale fiecărui lanț, este necesara

parcurgearea a doua margini ale grafului, prima în fata (urmărind un link ce pornește din nodul de unde s-a pornit drumul) și a doua înapoi către partea inițială din graf [6]. Ponderile autorităților sunt calculate că fiind distribuția lanțului ce verifica mai întăi un link direcționat către nodul de plecare, apoi urmandu-se un link direcționat către celelalte autorități ce se afla în legătură cu nodul respective. Ponderea hub-urilor este calculate similar, pornindu-se dintr-un nod aflat pe partea stangă a grafului și ajungând într-o autoritate de pe partea dreapta, de unde apoi se vă continua drumul pe marginile ce țintesc către autoritatea pe care s-a staționat în acel moment.

Astfel, se ia in considerare un lant Markov al parcurgerii nodurilor autoritate ale lui “G” si unul al parcurgerii nodurilor hub ale acestuia in scopul crearii scorului de autoritate “A(u)” si a scorului de hub “H(u)” (Fig.3) [3]:

**SALSA-Authority-Scores:**

1. Let $B^A$ be $\{u \in B : in(u) > 0\}$.
2. For all $u \in B$:
$$A(u) := \begin{cases} \frac{1}{|B^A|} & \text{if } u \in B^A \\ 0 & \text{otherwise} \end{cases}$$
3. Repeat until $A$ converges:
   (a) For all $u \in B^A$:
   $$A'(u) := \sum_{(v,u)\in N} \sum_{(v,w)\in N} \frac{A(w)}{out(v)in(w)}$$
   (b) For all $u \in B^A : A(u) := A'(u)$

**SALSA-Hub-Scores:**

1. Let $B^H$ be $\{u \in B : out(u) > 0\}$.
2. For all $u \in B$:
$$H(u) := \begin{cases} \frac{1}{|B^H|} & \text{if } u \in B^H \\ 0 & \text{otherwise} \end{cases}$$
3. Repeat until $H$ converges:
   (a) For all $u \in B^H$:
   $$H'(u) := \sum_{(u,v)\in N} \sum_{(w,v)\in N} \frac{H(w)}{in(v)out(w)}$$
   (b) For all $u \in B^H : H(u) := H'(u)$

*Fig. 3. Calculul scorului de autoritate si de hub al fiecarui nod din subgraful “A” si “H”*

Calculele realizate in Fig. 3. sunt reprezentate grafic in Fig.4:

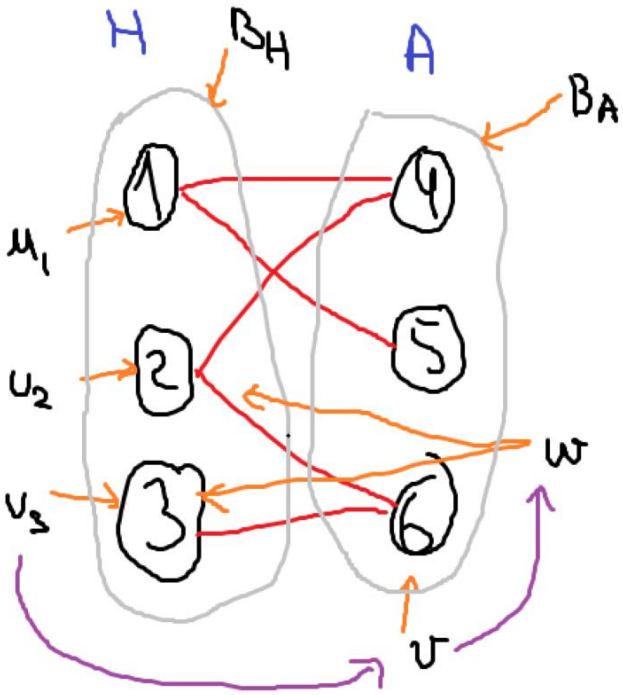


*Fig.4. Reprezentare grafica a drumurilor aleatorii realizate pentru “H”, respectic “A”*

Pentru calculul scorului de autoritate se vor allege intai toate nodurile din graf care au margini directionate de la “H” catre “A”. In cazul in care un nod nu va avea o margine directionata catre el, atunci scorul de autoritate al acestuia va fi 0. In caz contrar, acestuia I se va atribui o valoare egala cu raportul nodului fata de numarul total de noduri din subgraf.

În următoarea etapa, pentru toate nodurile aflate în “A” (notate cu “u”), se vă verifica fiecare nod din subgraful “H” care țintește către cel autoritate de la pasul curent (vertex-urile aflate în “H” se vor nota cu “v”), acest calcul reprezentând prima parte de drum aleatoriu. Al doilea pas este cel de verificare a nodurilor din “A” de care nodul “v” este legat (aceste noduri din “A” fiind notate cu “w”). Odată aflate toate traseele posibile, noua valoare de autoritate vă fi egala cu suma rapoartelor dintre valoarea autoritate a nodurilor “w” (la care un nod “u” poate ajunge prin trecerea către un nod “v” din subgraful “H” și înapoi printr-o legătură directa în partea “A”) și produsul dintre numărul marginilor ce ies din nodurile “v” cu cel al marginilor ce intra în nodurile “w”.

Aceasta etapa de calcul se vă repeta similar și pentru valoarea nodurilor hub, dar în sens inver (un drum începând din subgraful “H”, ce vă ajunge în “A “și se vă întoarce înapoi în zona inițială). Procesul de calcul pentru fiecare nod din graf se vă repeta de “k” ori, până când valoarea fiecărui nod se vă normaliza, ajungând că după mai multe iterații să nu se mai observe nicio diferență după recalculare.

**b) Metoda de implementare 2**

Că alternativa la calculul prezentat anterior, se poate crea inițial o matrice de adiacenta “W” a grafului direcționat “C”. Prin intermediul acesteia, se vor crea încă 2 matrice, unde “Wr” vă fi matricea rezultata prin împărțirea fiecărei intrări diferite de 0 la suma elementelor de pe rânduri, respective “Wc”, o matrice rezultata din împărțirea tuturor elementelor nenule ale celei inițiale la suma elementelor de pe coloana.

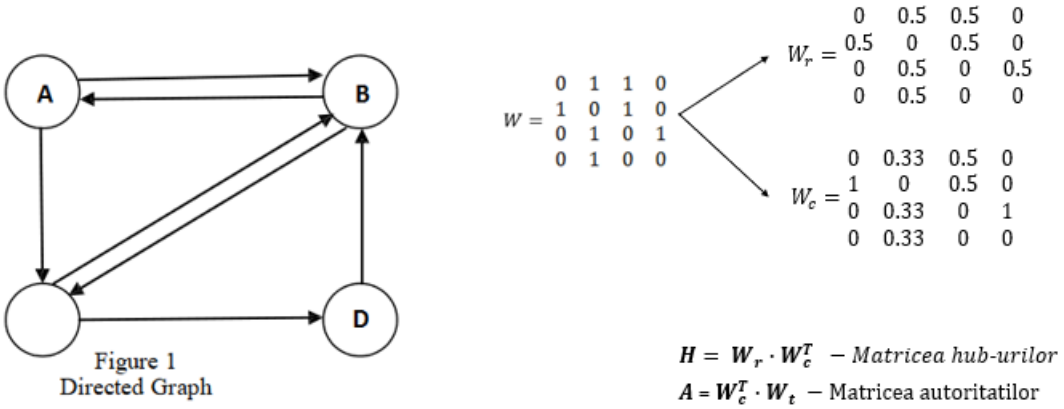


Fig.5. Crearea matricei de adiacenta in functie de graful initial

Dupa calculul acestor matrice, principalele comunitati de autoritati si hub-uri vor fi constituite din primele “k” pagini avand ca cele mai mari intrari in vectorul principal al A si H.

In cadrul implementarii acestui algoritm, s-a mers pe prima metoda de implementare, fiind mai rapid de calculat prin intermediul librariei “networkx” din Python.

## 3. Rezultate și probleme cunoscute

Unul dintre avantajele acestui algoritm în comparație cu celelalte metode utilizate în web structure mining este acela al lipsei sensibilității acute la efectul TKC (Tightly Knit Community), o problema majora pentru algoritmii de web mining ce nu pot să ofere notații de importanta precise grupurilor strânse, unde mai multe elemente legate de același subiect se leagă intre ele, formând un

amalgam de trimiteri ce își consolidează reciproc importanta, fiecare pagina având multe link-uri tip hub și multe referințe tip autoritate).

Un alt avantaj al acestui algoritm este acela al rezultatelor remarcabile atunci când este supus la căutări generale, având rezistenta la hub-urile cu o valoare foarte mica, astfel crescând șansă găsirii rezultatelor relevante. De asemenea, un avantaj similar algoritmului PageRank este acela că ponderile nodurilor pot fi puse explicit, putând astfel influenta într-o anumită măsură căutările inițiale, în funcție de scopul dorit (căutări generale, căutări de informații restrânse, etc.). În plus, spre deosebire de celelalte metode de web structure mining, SALSA nu are topic drift, astfel prezentând un avantaj în cazul problemelor de polisemie.

Printre dezavantajele cunoscute ale acestui algoritm este acela al favorizării în unele cazurilor a comunităților mai mici, fară, poate, prea mare legătură cu subiectul (ce cauzează automat mai puține noduri, ajungând că rezultatele să se normalizeze mult mai repede, data de cazurile cu eșantioane mari de date).

Un alt dezavantaj al acestui algoritm este reprezentat de rezultate mai slabe decât HITS sau PageRank în cazul căutărilor specific, neavand capabilitatea de a normaliza rezultatele în urma mai multor iterații pe un număr mic de noduri. În plus, acest algoritm este dependent de interogări, astfel putând avea rezultate maim ult sau mai puțin bune în funcție de căutările inițiale (graficul inițial “C” având un impact major asupra rezultatelor extrapolate din graficul “G”).

## 3. Seturi de date utilizare (nume si eșantioane)

In cadrul acestui proiect au fost utilizate 4 esantioane de date privind relatiile dintre persoane de pe diferite retele sociale, obtinute de pe site-ul https://snap.stanford.edu/data/index.html.

Seturile de date sunt:

- Twitter (“Social circles: Twitter”): Dataset ce contine cercuri de prieteni de pe Twitter luate din surse publice. Dataset-ul formeaza un graf directionat constituit din 81.306 noduri cu 1.768.149 muchii;
- Google Plus (“Social circles: Google+”): Dataset ce contine cercuri de prieteni de pe Google+, informatiile find colectate de la oameni care si-au distribuit relatiile cu ceilalti membri. Dataset-ul formeaza un graf directionat constituit din 107.614 noduri si 13.673.453 muchii;
- Facebook(“Social circles: Facebook): Dataset ce contine listele de prieteni de pe Facebook, informatiile fiind colectate de la participantii unor studii pe “Facebook app”. Dataset-ul formeaza un graf nedirectionat constituit din 4.039 noduri si 88.234 muchii;
- GitHub(“GitHub Social Network”): Dataset ce contine oameni activi pe GitHub care se urmaresc intre ei. Dataset-ul formeaza un graf nedirectionat constituit din 37.700 noduri si 289.003 muchii.

**Pentru** fiecare din esantioanele prezentate anterior s-au realizat verificari pe diferite esantioane: 100, 1000, 10.000, 100.000 si intregul dataset.

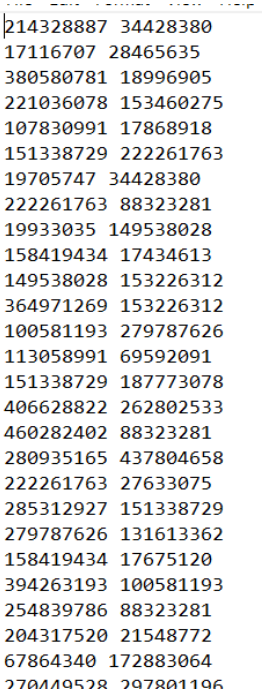


*Fig.6. Exemplu de eșantion din setul "Twitter"*

## 4. Rezultate și evaluarea acestora

In urma rularii programului pe un set de date format din cercuri sociale de prieteni de pe platforma sociala "Twitter", s-a observant ca dupa 10 iteratii ale algoritmului asupra grafului, atat valorile autoritatilor cat si ale grafurilor s-au normalizat spre o singura valoare unitara. Din aceste calculi se poate observa ca atenuarea valorilor autoritatilor si hub-urilor este realizata rapid din punct de vedere al iteratiilor, fiind foarte efficient in gasirea imediata a raspunsurilor dorite.

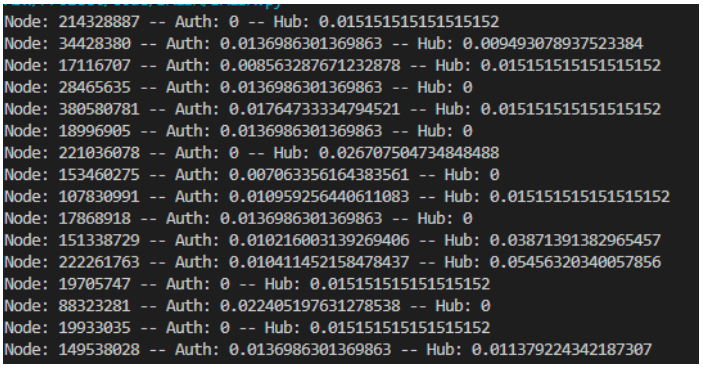


*Fig.7. Rezultatele dupa 5 iteratii*

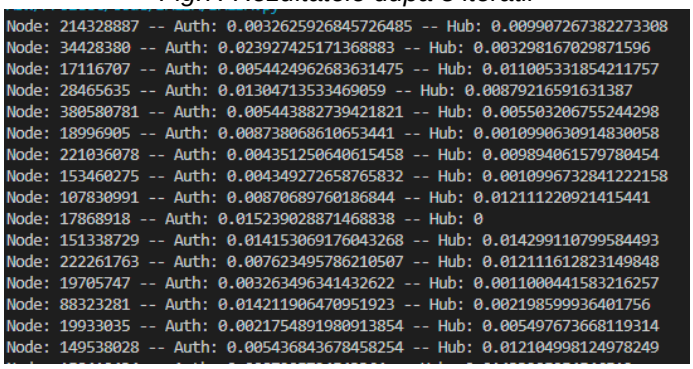


*Fig.8. Rezultatele dupa 10 iteratii*

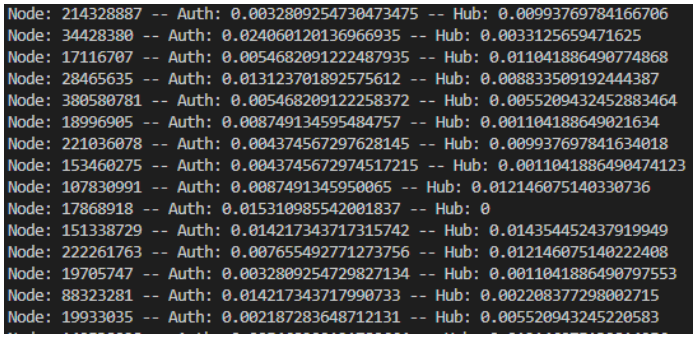


*Fig.9. Rezultatele dupa 100 iteratii*

In urma rularii, s-a constatat problema depistata la punctul 3, aceea ca algoritmul SALSA este dependent de interogari (si mai ales de numarul acestora), valorile multor noduri devenind 0, atat la autoritati cat si la hub-uri, din cauza lipsei unor valori catre care vectorii sa fie indreptati. In acest sens, cu cat marimea esantionului este mai mica, cu atat algoritmul este mai putin precis (lucru care se poate aplica, de asemenea, si la cautarile specifice unde informatia nu este atat de abundenta (Fig.9,Fig.10,Fig.11).

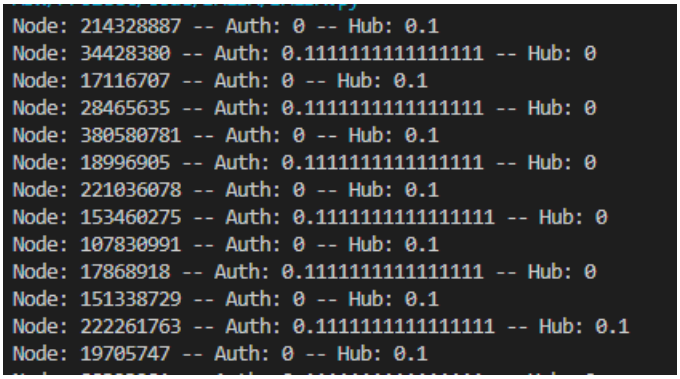


Fig.10. Cautari pe un esantion de 10 noduri din set

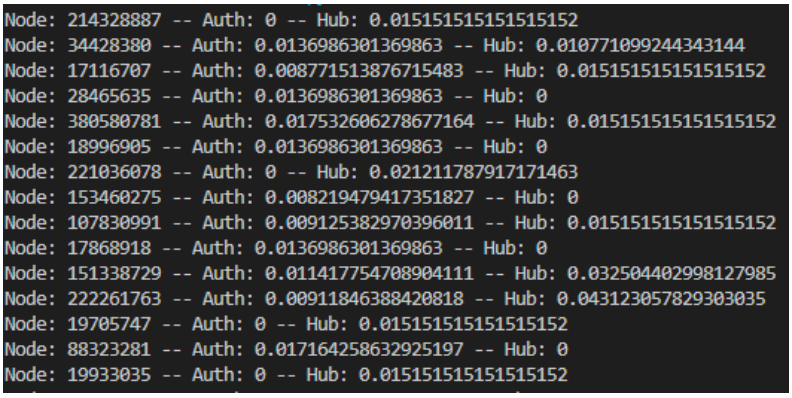


Fig.11. Cautari pe un esantion de 100 noduri din set

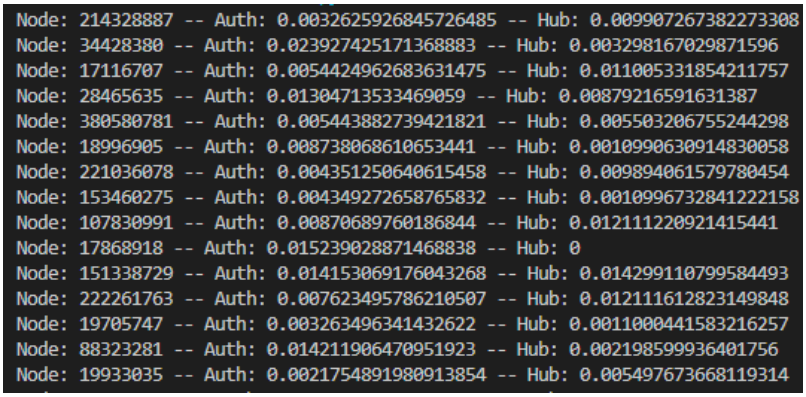


Fig.12. Cautari pe un esantion de 1000 noduri din set

O alta observatie in urma diferitelor rulari este aceea a implementarii tipurilor de grafuri, lucru ce influenteaza semnificativ rezultatele finale. In cadrul implementarii s-a considerat realizarea unui graf directionat pentru a putea obtine mai usor marginile care intra si ies din diferite noduri, insa la verificarea setului de date cu grafuri nedirectionate, numarul valorilor hub-urilor egale cu 0 au crescut considerabil. Acest lucru, insa, va putea fi remediat prin intermediul unei implementari diferite a programului.

Pe langa aceasta implementare, s-a constatat impactul major pe care ponderea initiala a valorii A(u) (respective H(u)) o are asupra rezultatului final, autoritatilor indreptandu-se spre valoari unitare in cazul echivalarii initiale a autoritatilor si hub-urilor cu 1.

## 5. Bibliografie

[1] R. Lempel and S. Moran. 2001. SALSA: the stochastic approach for link-structure analysis. ACM Trans. Inf. Syst. 19, 2 (April 2001), 131–160. https://doi.org/10.1145/382979.383041

[2] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas. 2005. Link analysis ranking: algorithms, theory, and experiments. ACM Trans. Internet Technol. 5, 1 (February 2005), 231–297. https://doi.org/10.1145/1052934.1052942

[3] Najork, Marc. (2007). Comparing the effectiveness of hits and salsa. 157-164. 10.1145/1321440.1321465

[4] Marc A. Najork. 2007. Comparing the effectiveness of hits and salsa. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM '07). Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/10.1145/1321440.1321465

[5] Farahat, Ayman, Thomas LoFaro, Joel C. Miller, Gregory Rae, and Lesley A. Ward. “Authority Rankings from HITS, PageRank, and SALSA: Existence, Uniqueness, and Effect of Initialization.” SIAM Journal on Scientific Computing 27, no. 4 (2006): 1181–1201. doi:10.1137/S1064827502412875

[6] Mellah, Mohamed & Amine, Abdelmalek & Hamou, Reda & Kumar, A.V.. (2014). Link Analysis for Communities Detection on Facebook. International Journal of Data Mining And Emerging Technologies. 4. 10.5958/2249-3220.2014.00017.2.

[7] http://pubs.sciepub.com/ajss/3/2/3/index.html, accesat: 22.05.2022, 14:00

[8] https://docs.oracle.com/cd/E56133_01/2.4.0/reference/algorithms/salsa.html, accesat: 22.05.2022, 14:00

[9] https://snap.stanford.edu/data/index.html#web, accesat: 23.05.2022, 14:00

[10] https://slideplayer.com/slide/6060463/, accesat: 22.05.2022, 14:00