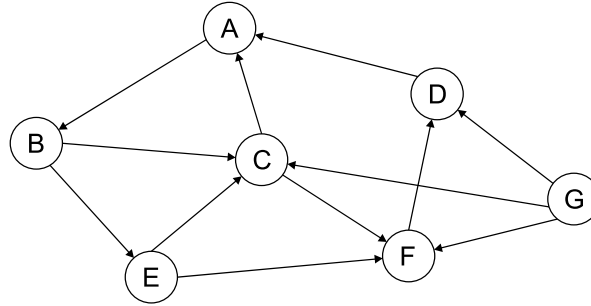


Operativni sistemi 2 – I popravni kolokvijum – hadoop
Institut za matematiku i informatiku, PMF Kragujevac
22. januar. 2016. godine

Na slici ispod (Slika 1) je prikazan usmereni graf sa stranicama na web-u i njihovim vezama.



Slika 1- Linkovi i veze između njih

Linkovi i veze između njih su predstavljeni u fajlovima na *HDFS* fajl sistemu sa sledećom strukturom:

Naziv_čvora_n -> Link_do_cvora1, Link_do_cvora2, ..., Link_do_cvora_m

Za primer sa Slike 1 fajl bi izgledao kao u nastavku:

A -> B
B -> C, E
C -> A, F
D -> A
E -> C, F
F -> D
G -> C, D, F

Potrebno je izračunati važnost i pouzdanost svakog od linka uz pomoć *PageRank* algoritma. Iterativni postupak rada *PageRank* algoritma je dat u nastavku:

$$\text{I)} \quad PR_0(p_i) = \frac{1}{N},$$

$$\text{II)} \quad PR_{k+1}(p_i) = \frac{1-d}{N} + d * \sum_{j=1}^{m_i} \frac{PR_k(p_j)}{L(p_j)}$$

$PR_{k+1}(p_i)$ - PageRank stranice p_i u $k+1$ koraku,

$PR_k(p_j)$ - PageRank stranice j u k -tom koraku koja ima link ka stranici p_i ,

$L(p_j)$ – Broj linkova stranice j ka ostalim stranicama u grafu,

$d = 0.85$,

N - Ukupan broj linkova u mreži.

U nastavku je dat primer računanja vrednosti *PageRank*-a za stranicu **A** u $k+1$ koraku:

$$PR_{k+1}(A) = \frac{1-d}{N} + d * \left(\frac{PR_k(C)}{2} + \frac{PR_k(D)}{1} \right)$$

Programsko rešenje treba da sadrži odgovarajuće klase sa konfiguracijom za pokretanje posla na *Hadoop-u* kao i bash skriptu koja kompajlira odgovarajuće fajlove i pokreće posao. Bash skripta mora imati naziv **pokreni.sh** i prihvatati 3 ulazna parametra:

1. Putanja do ulaznih fajlova na *HDFS* fajl sistemu,
2. Putanja do foldera gde ce se naći izlazni fajlovi,
3. Broj čvorova u grafu,
4. Broj iterativnih koraka u postupku računanja *PageRank-a* (maksimalna vrednost *k*)