Wyszukiwanie informacji w sieci, Lista 6 **1.06.2007**

1. [1] Rozważmy graf $G = (V, E)$ z wierzchołkami $V = \{1, 2, 3, 4\}$ i krawędziami $E = \{(1, 2), (2, 3), (3, 4), (4, 2)\}$. Niech prawdopodobieństwo teleportacji wynosi $\alpha = 0.01$. Podaj macierz A procesu Markowa dla metody PageRank. Wyznacz PageRank:

   (a) przybliżony: iteracyjnie, wykonując cztery kroki, startując z wektora $r = (0.25, 0.25, 0.25, 0.25)$;

   (b) dokładny: poprzez wyznaczenie wektora własnego bezpośrednio, tzn. znajdując rozwiązania odpowiedniego układu równań liniowych (podaj ten układ równań).

   Na ćwiczeniach przedstaw *wyniki* obliczeń.

2. [0.5] (21.12) Rozważmy algorytm PageRank z prawdopodobieństwem teleportacji równym α, gdzie n to liczba wierzchołków grafu. Pokaż, że pagerank każdej strony wynosi co najmniej α/n. W oparciu o ten fakt wyciągnij wnioski na temat różnic między rangami różnych dokumentów gdy α jest bliskie 1.

3. [1.5] Przyjmijmy, że prawdopodobieństwo teleportacji wynosi α.

   (a) Opisz jak zmienia się w iteracyjnej metodzie wyznaczania PageRank ranga wierzchołka, który w oryginalnym grafie ma stopień wejściowy 0.

   (b) Opisz jaki wpływ na ostateczne wartości PageRank mają wierzchołki o stopniu wyjściowym 0.

   (c) Opisz wpływ wierzchołków o stopniu wejściowym/wyjściowym zero na wyniki algorytmu HITS.

4. [1] Niech $x_1, \ldots, x_n \in R^n$ będą wektorami własnymi macierzy $A \in R^{n \times n}$, a $\lambda_1, \ldots, \lambda_k \in R$ odpowiadającymi im wartościami własnymi. Niech też $1 = |\lambda_1| > \ldots > |\lambda_n|$. Załóżmy, że każdy wektor $v \in R^n$ można jednoznacznie przedstawić jako kombinację liniową wektorów $x_1, \ldots, x_n$. Pokaż, że ciąg $Av, Av^2, Av^3, \ldots$ zbiega do wartości x_1 dla każdego $v \neq 0^n$.

5. [1] (21.13-21.15) W metodzie PageRank zorientowanej na temat przyjmuje się, że algorytm PageRank uruchamiamy dla grafu Y będącego podgrafem całej sieci, zawierającym podgraf S stron dotyczących danego tematu.

   (a) Załóżmy, że $Y = S$. Jak wówczas zdefiniujesz macierz odpowiedniego procesu Markowa? Pamiętaj, że otrzymany w efekcie proces Markowa powinien być ergodyczny.

   (b) Uzasadnij, że wybór Y będącego pewnym właściwym nadgrafem S może dać lepsze wyniki niż $Y = S$. Zaproponuj taki wybór Y.

   (c) Niech s będzie stroną należącą do klasy S. Czy wartość pagerank s w klasie K musi być większa (bądź równa) wartości pagerank strony s w całej kolekcji dokumentów?

6. [1] (21.16-21.17) Rozważmy sytuację, w której mamy wyliczone wartości pagerank dla stron w obrębie dwóch kategorii K_1 i K_2, przy tym samym prawdopodobieństwie teleportacji równym α. Rozważmy teraz użytkownika zainteresowanego zarówno kategorią K_1 jak i K_2, ale w proporcjach takich, że K_1 preferuje w $q \cdot 100\%$ dla $0 < q < 1$. Załóżmy więc, że w przypadku teleportacji (wybieranej z prawdopodobieństwem α) skaczemy do stron z kategorii K_1 z prawdopodobieństwem q a do stron z kategorii K_2 z prawdopodobieństwem $(1-q)$. Oczywiście, teleportacja do każdej ze stron w obrębie danej kategorii jest tak samo prawdopodobna.

   Dla uproszczenia możesz założyć, że

   - pagerank w obrębie kategorii liczymy tylko dla podgrafu stron z tej kategorii (w szczególności, dla $K_1 \cup K_2$ rozważamy sumę podgrafu odpowiadającego K_1 i podgrafu odpowiadającego K_2);
   - nie ma linków między K_1 i K_2.

   Twoje zadanie:

   (a) Opisz macierz procesu Markowa odpowiadającego rankingowi dla $K_1 \cup K_2$, uzależniając ją od macierzy odpowiadających procesom Markowa dla K_1 i K_2 oraz od q.

   (b) Pokaż, że zdefiniowany w tym zadaniu proces Markowa jest ergodyczny. A zatem, rozkład stacjonarny można wyznaczyć jako wektor własny macierzy opisanej w poprzednim podpunkcie.

7. [1] (21.18) Dla sytuacji opisanej w poprzednim zadaniu uzasadnij, że pagerank dla kategorii $K_1 \cup K_2$ z parametrem q można uzyskać jako kombinację liniową $qp_{K_1} + (1-q)p_{K_2}$, gdzie p_{K_i} to wektor pagerank dla kategorii K_i (wartości rang wierzchołków nie należących do K_i są równe 0).

8. [0.5] Algorytm HITS wykonuje obliczenia na małym podgrafie całej sieci. Przedstaw sposób wyboru podgrafu i uzasadnienie dla takiego wyboru. (Oprzyj się na oryginalnym artykule Jona Kleinberga).