:Search Engines Algorithms

1. Google's PageRank

3. Hubs & Authorities (KLEINBERG)

PageRank

- PageRank הוא ערך מספרי אשר מיצג את מידת החשיבות של דף ב-web.
- Google מניח כי כאשר דף מקשר (links) לדף אחר הוא בעצם "מצביע" לטובת הדף האחר.
- ככל שיש יותר הצבעות לדף מסוים - הדף בעל חשיבות גבוהה יותר.
- כמו כן, מידת החשיבות של הדף שמבצע את ההצבעה קובעת כמה ההצבעה עצמה חשובה.

PageRank – cont’

- Google מחשב את מידת החשיבות של דף על סמך ההצבעות עליו ועל סמך מידת החשיבות של כל הצבעה.

- ה- PageRank של הדפים שמקשרים אל דף מסוים נלקח בחשבון אך באופן פרופורציוני לסך כל הדפים אליהם מצביע הדף.

הערה: לא כל הקישורים נלקחים בחשבון ע"י google.

?How is PageRank calculated

$$PR(A) = (1 \quad d) + d\left(\frac{PR(t_1)}{C(t_1)} + \cdots + \frac{PR(t_n)}{C(t_n)}\right)$$

$t_1, \ldots t_n$ *pages linking to page A*

$C(t_i)$ *- number of outbound links that page* t_i *has*

d *- damping factor* $(usually\ d = 0.85)$

PageRank – cont’

בעיה: נניח כי קיימים 2 דפים A, B המקשרים האחד אל השני בלבד (אין להם קשרים נוספים מכל סוג שהוא).

אזי:

- נחשב את ה-PageRank של דף A.
- לדף A יש PageRank חדש אך החישוב כלל את הערך של דף B המצביע לדף A.
- כמו כן, דף B מוצבע ע"י דף A, לדף A יש ערך PageRank חדש שלא נכלל בחישוב ה- PageRank של B ← ערך ה- PageRank החדש של A מבוסס על מידע שאינו מדויק ולכן אינו מדויק..
- כנ"ל לגבי חישוב ה- PageRank של דף B.

PageRank – cont'

! לא ניתן לחשב את ה- PageRank של דף A עד אשר נדע את ה- PageRank של דף B וכמו כן – לא ניתן לחשב את ה- PageRank של דף B לפני שנדע את ה- PageRank של דף A.

- ניתן להתגבר על הבעיה ע"י ביצוע החישובים פעמים רבות כאשר בכל פעם נשתמש בערכי ה- PageRank העדכניים ביותר.
- ידרשו 40-50 איטרציות עד להתכנסות של הערכים.

שאלה 1

נתונים 3 דפי A, B, C: web עם הקשרים הבאים:

- A מקשר אל B ואל C.
- B מקשר אל C.
- C מקשר אל A.

מצאו את ה-PageRank של כל דף. הניחו כי damping factor=0.

שאלה 2

נתונים 3 דפי A, B, C: web עם הקשרים הבאים:

- A מקשר אל B ואל C.
- B מקשר אל C.
- C לא מקשר לאף דף.

מצאו את ה-PageRank של כל דף. הניחו כי damping factor=30%.

Hubs & Authorities (KLEINBERG)

- האלגוריתם מתבסס על הקשר הקיים בין סמכויות לנושא מסוים (authorities) לבין הדפים המקשרים לסמכויות רבות הקשורות לנושא (hubs).

- האלגוריתם מופעל על תת-גרף ממוקד של ה-WWW אשר נבנה על-סמך פלט של מנועי חיפוש מבוססי טקסט. הטכניקה לבניית תת-הגרף תוכננה ליצירת אוספים קטנים של דפים האמורים להכיל את הדפים ה"סמכותיים" ביותר לנושא המבוקש.

בניית תת-גרף ממוקד של ה-WWW

- נסתכל על אוסף V כלשהו של hyperlinked pages כעל גרף מכוון G=(V,E). צמתי הגרף הם הדפים וקשת מכוונת (p,q) ב-E מציינת קיום קישור (link) מ-p אל q.
- מגרף G ניתן לבנות תת-גרף: אם $W \subseteq V$ הוא תת קבוצה של דפים נסמן ב- G[W] את הגרף המושרה על W. צמתי הגרף הם דפים ב-W, קשתות הגרף הן כל הקשרים בין דפים ב-W בהתאמה.
- אלגוריתם Subgraph(σ,s,t,d) מחזיר קבוצת דפים S_σ אשר מהווה את קבוצת הבסיס עבור המחרוזת σ.
- $G[S_\sigma]$ הוא תת-הגרף המושרה על הדפים ב- S_σ .

$\underline{Subgraph(\sigma, s, t, d)}$

σ: a query string.

s: a text-based search engine.

t,d: natural numbers.

Let R_σ denote the top t results of s on σ.

Set $S_\sigma := R_\sigma$

For each page $p \in R_\sigma$

Let $\Gamma^+(p)$ denote the set of all pages p points to.

Let $\Gamma^-(p)$ denote the set of all pages pointing to p.

Add all pages in $\Gamma^+(p)$ to S_σ.

If $|\Gamma^-(p)| \leq d$, then

Add all pages in $\Gamma^-(p)$ to S_σ.

Else

Add an arbitrary set of d pages from $\Gamma^-(p)$ to S_σ.

End

Return S_σ

האלגוריתם האיטרטיבי

- The mutually reinforcing relationship:

  A good hub is a page that points to many good authorities, a good authority is a page that is pointed by many good hubs.

- Page's **authority weight**: $x^{\langle p \rangle}$.
- Page's **hub weight**: $y^{\langle p \rangle}$.

$\underline{Iterate(G,K)}$

G: *a collection of n linked pages.*

k: *a natural number.*

Let z denote the vector $(1,1,1,\ldots,1)\in R^n$.

Set $x_0 := z$.

Set $y_0 := z$.

For $i=1$ *to* k

Apply the O_1 *operation to* (x_{i-1}, y_{i-1}), *obtaining new x-weights* x'_i.

Apply the O_2 *operation to* (x'_i, y_{i-1}), *obtaining new y-weights* y'_i.

Normalize x'_i *obtaining* x_i.

Normalize y'_i *obtaining* y_i.

End

Return (x_k, y_k).

$$\text{Operation } O_1:\ x^{\langle p\rangle} \leftarrow \sum_{\{q:(q,p)\in E\}} y^{\langle p\rangle}$$

$$\text{Operation } O_2:\ y^{\langle p\rangle} \leftarrow \sum_{\{q:(p,q)\in E\}} x^{\langle p\rangle}$$

$$\text{Normalize } 'x_i:\ \sum_{p\in S_\sigma} (x^{\langle p\rangle})^2 = 1$$

$$\text{Normalize } 'y_i:\ \sum_{p\in S_\sigma} (y^{\langle p\rangle})^2 = 1$$

סינון

- על מנת לסנן את c הסמכויות הטובות ביותר ו-c המרכזים הטובים ביותר נשתמש בפרוצדורה הבאה:

$\underline{Filter(G,k,c)}$

G: *a collection of n linked pages.*

k,c: *natural numbers.*

$(x_k, y_k) := Iterate(G,k)$.

Report the pages with the c largest coordinates in x_k *as authorities.*

Report the pages with the c largest coordinates in y_k *as hubs.*

הגרסא המתמטית של האלגוריתם

- בהינתן גרף G=(V,E) כאשר $V=\{p_1, p_2, \ldots, p_n\}$ תהי A מטריצת הסמיכויות של הגרף G. בכניסה ה- (i, j) ב-A יהיה ערך 1 אם (p_i, p_j) קשת ב-G ו-0 אחרת.

- h הוא ווקטור של מידת ה-hubbiness של כל דף ו- a הוא ווקטור של מידת ה- authority של כל דף.

- נחשב את h ו- a באופן הבא:

$$h = AA^T h$$
$$a = A^T Aa$$

שאלה

נתונים 3 דפי A, B, C: web עם הקשרים הבאים:

- A מקשר אל B ואל C.
- B מקשר אל C.
- C מקשר לעצמו.

מצאו את מידת ה-hubbiness וה-authority של כל דף.