בס״ד

עבודה סמינריונית בנושא

Specral Graph Theory

בהנחיית: פרופ׳ אוריאל פייגה

מגישה: אילנה ברזניק

אוקטובר 2012

תוכן עניינים:

מבוא

מטרתי בעבודה זו היא להציג את אחת השיטות מהשנים האחרונות לדרוג הדפים באינטרנט. השיטה הוצגה על ידי קליינברג במאמרו JON M. KEINBERG/Authoritative Sources in a Hyperlinked Environment/*Journal of the ACM, Vol. 46, No. 5, September 1999, pp. 604-632*. טענתו המרכזית של קליינברג היא כי ניתן ללמוד הרבה על תוכנה של הסביבה המתוקשרת בעזרת ניתוח המבנה שלה. תנאי הכרחי ללמידה זו הוא הימצאותם של הכלים המתאימים. בעבודתו המוצגת במאמר מציג קליינברג כלים אלגוריתמיים שמטרתם לנתח את מבנה האינטרנט ואת הקשרים בו ככלי לקבלת אינפורמציה על תוכנו.

באמצע שנות ה-90 היה Yahoo האתר הראשון שדרג את הדפים באינטרנט בצורה ידנית. כיום, האינטרנט כבר גדול מידי ולכן שיטה זו הופכת להיות יקרה ולא אפקטיבית. השאלה הגדולה היא כיצד לעשות את הדרוג בצורה אוטומטית?. אפשרות אחת היא להמציא תוכנה הקוראת את דפי האינטרנט ומדרגת אותם לפי איכותם. כיום תוכנה כזו טרם הומצאה עקב המורכבות שבעניין. לא נפסלת האפשרות כי תוכנה כזו תומצא בעתיד. אפשרות אחרת היא לחשוב על מדד שייתן אינדיקציה על איכות הדפים באינטרנט ותוך כדי שהדבר יוכל להיעשות בצורה יחסית אוטומטית על ידי מחשב.

עבודתו של קליינברג היא מהראשונות המציעות איך לעשות את דרוג הדפים לא רק לפי תוכנם אלא על ידי בחינת דפי האינטרנט המצביעים עליהם, שהם מהווים כמעין "ממליצים". לצורך בדיקת איכותו של דף אינטרנט לא מספיק לבדוק את כמות הממליצים. שכן, אם זהו הקריטריון לבחינת חשיבותו של דף האינטנרט ניתן לבנות בצורה מלאכותית הרבה דפי אינטרנט שיצביעו על דף זה. דרך נוספת שניתן לדרג בה את הדפים היא לפי מספר הפעמים בהם מופיעה מילת החיפוש בדף. לדוגמא, אנו רוצים אינפורמציה באינטרנט על חלב. נבצע חיפוש באינטרנט ומנוע החיפוש יחזיר לנו דפים המדורגים לפי מספר הפעמים שמופיעה בהם המילה "חלב". גם שיטה זו אינה אפקטיבית במידה מספקת שכן אם העוסקים ב"קידום אתרים" מגלים אותה הם יוכלו לקחת דף אינטרנט שברצונם לקדם ולקדם **אותו** על ידי הוספת המילה המתאימה מספר רב של פעמים בתוכן הדף. בהמשך העבודה נראה מה מציע לעשות קליינברג.

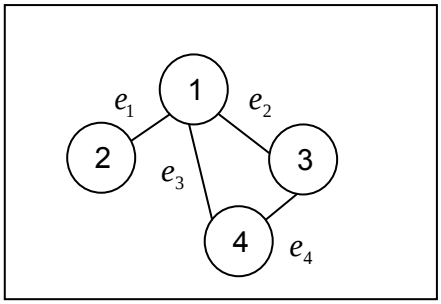
מהו גרף

גרף הינו ייצוג מופשט של קבוצה של אובייקטים, כאשר כל תת-קבוצה של אובייקטים בקבוצה עשויים להיות מקושרים זה לזה. האובייקטים הניתנים לקישור מכונים קודקודים או צמתים (באנגלית: vertex), וקבוצת הקודקודים מסומנת באות V.

הקישורים בין הקודקודים מכונים צלעות או קשתות (באנגלית: edge), וקבוצת הצלעות מסומנת באות E. מתקיים כי קבוצת הצלעות מקיימת: $E \subseteq V \times V$, כלומר: כל צלע הינה זוג הקודקודים, אותם היא מקשרת.

גרף, אשר קבוצת הקודקודים שלו היא V וקבוצת הצלעות שלו היא E מסומן באופן הבא: $G = (V, E)$.

לדוגמא, נתבונן בגרף הבא:



תורת הגרפים הספקטלרית היא תורה המאפשרת להתאים לכל גרף מטריצה וללמוד מהערכים העצמיים של המטריצה את תכונות הגרף.

הצגת גרף בעזרת מטריצה

ניתן לייצג גרף בדוגמא על ידי כמה מטריצות:

1. מטריצת קודקודים-קשתות:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} e_1 & e_2 & e_3 & e_4 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

מטריצה זו מוגדרת על הקודקודים והקשתות על ידי ההגדרה הבאה:

$$a_{i,j} = \begin{cases} 1 & \text{אם קודקוד j עובר דרך צלע i} \\ 0 & \text{אחרת} \end{cases}$$

2. מטריצת הסמיכויות:

מוגדרת על הקודקודים.

$$a_{i,j} = \begin{cases} 1 & \text{אם } v_i \text{ סמוך ל } v_j \\ 0 & \text{אחרת} \end{cases}$$

בדוגמא שלנו המטריצה היא:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

3. מטריצת הלפלסיאן:

מטריצה זו מוגדרת על הקודקודים בלבד.

$$l_{i,j} = \begin{cases} \deg(v_i) & \text{אם } i = j \\ -1 & \text{אם } i \neq j \text{ וגם } v_i \text{ סמוך ל } v_j \\ 0 & \text{אחרת} \end{cases}$$

הצורה המנורמלת של מטריצה זו היא :

$$l_{i,j} = \begin{cases} 1 & \text{אם } i = j \text{ וגם } v_i \neq 0 \\ -\frac{1}{\sqrt{\deg(v_i) \cdot \deg(v_j)}} & \text{אם } i \neq j \text{ וגם } v_i \text{ סמוך ל } v_j \\ 0 & \text{אחרת} \end{cases}$$

כאשר הגרף הוא רגולרי, כלומר הדרגות של כל הקודקודים שוות, שתי המטריצות שקולות.

נבנה את המטריצות עבור הגרף שבדוגמא.

המטריצה הרגילה :

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix} \end{array}$$

המטריצה המנורמלת :

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 1 & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{3}} & 1 & 0 & 0 \\ -\frac{1}{\sqrt{6}} & 0 & 1 & \frac{1}{2} \\ -\frac{1}{\sqrt{6}} & 0 & \frac{1}{2} & 1 \end{pmatrix} \end{array}$$

2

3

4

נשים לב שבגרפים לא מכוונים המטריצות הנ״ל הן סימטריות.

צורה נוספת להסתכל על מטריצת הלפלסיאן היא להסתכל על מטריצה המתקבלת מההפרש בין מטריצת הדרגות למטריצת הסמיכויות של גרף G. על מטריצת הדרגות היא מטריצה סקלרית שבה מופיעות על האלכסון דרגות הקודקודים.

לדוגמא, בגרף הקודם:

סימון:

$$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix} - \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix}$$

$$L_G = D_G - A_G$$

הערכים העצמיים והוקטורים העצמיים של מטריצות אלו מספקים אינפורמציה על הגרף. עבור גרפים רגולריים, המידע שניתן להסיק מאחת ההצגות של המטריצה זהה למידע שניתן להסיק מההצגות האחרות. לעומת זאת, בגרפים שאינם רגולריים, בחירת ההצגה היא דבר המשנה מאוד.

הסבר לכך שבגרף רגולרי הייצוג לא משנה: ניתן לעשות טרנספורמציה פשוטה בין הערכים העצמיים של מטריצת הסמיכויות לערכים העצמיים של מטריצת הלפלסיאן.

במקרה של גרף d רגולרי הטרנספורמציה היא: $\lambda \to d - \lambda$

מטריצת הסמיכויות $Lap = d \cdot I -$

⇓ ⇓

העי״ע מוכפל ב 1 - נוסף d לעי״ע

הוי״ע לא משתנה הוי״ע לא משתנה

הערה: לאורך כל העבודה נשתמש במשפט מאלגברה לינארית שאם המטריצה היא ממשית וסימטרית אז הערכים העצמיים שלה ממשיים, והווקטורים העצמיים מהווים בסיס ל R^n. כלומר הם פורסים ובלתי תלויים.

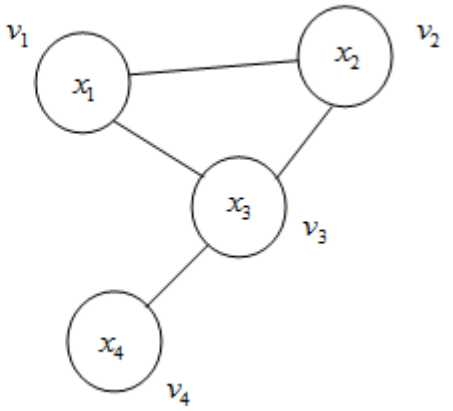
אינטרפרטציה חדשה למושג הערך העצמי במטריצות

אנו רגילים מאלגברה לינארית כי עבור מטריצה A , x הוא וקטור עצמי עם ערך עצמי λ , אם מתקיים $Ax = \lambda x$.

נסתכל על אינטרפרטציה חדשה למושג הערך העצמי,

הקודקוד.

המשקל=הערך שלו בוקטור העצמי.



$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} & & & \\ & & & \\ & & & \\ & & & \end{pmatrix} \end{array}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

⇩

וקטור עצמי

λx_1 = סכום המשקלים של הקודקוד v_1 .

לדוגמא: מסכמים את המשקל של השכנים (המחוברים אליו בקשת) של v_3 ומקבלים $\lambda \cdot v_3$.

רקע תיאורטי Spectral Graph Theory

משפט 1: 0 הוא תמיד ערך עצמי של מטריצת הלפלסיאן .

הוכחה: (1,1...,1,1) הוא תמיד וקטור עצמי כי כל רכיב בוקטור Ax הוא למעשה סכום האיברים בכל שורה במטריצה A . הדבר קורה מכיוון ש x הוא וקטור האחדים. סכום זה הוא בדיוק 0 כי הוא מורכב מ:

הדרגה של כל קודקוד+(1-)*מס' הקודקודים הסמוכים לו=0.

משפט 2: מספר הפעמים שבהם 0 הוא ערך עצמי של מטריצת הלפלסיאן הוא כמספר רכיבי הקשירות של הגרף.

הוכחה: נניח שלגרף G, K רכיבי קשירות בגודל $N_1,...,N_k$. נחלק את המטריצה לפי רכיבי קשירות. נסדר את המטריצה כך ש N_1 השורות הראשונות של L יהיו שייכות לרכיב הראשון. N_2 השורות הבאות יהיו שייכות לרכיב השני וכך הלאה.... נקבל מטריצה מהצורה:

$$L=\begin{pmatrix} L_{1,1} & 0 & \cdots & 0 \\ 0 & L_{2,2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & L_{k,k} \end{pmatrix}$$

לפי משפט באלגברה לינארית: אם מחליפים את סדר השורות ובאותו האופן את סדר העמודות הערכים העצמיים לא משתנים כי פשוט שינינו את סדר הקודקודים במטריצה.

המטריצה $L_{i,i}$ היא תת-מטריצה $N_i \times N_i$ המייצגת את מטריצת הלפלסיאן של הרכיב ה- i של G.

נגדיר את הוקטור x_i כוקטור עם קואורדינטות 1 בכל הקואורדינאטות המתאימות לרכיב ה- i ו 0 בשאר הקואורדינאטות. נעשה זאת עבור i=1,...,k .

כל הוקטורים $x_1, x_2, \ldots$ הם בת"ל כי המכפלה הפנימית של כל שניים מהם היא 0 מה שאומר שהם ניצבים. לכן יש ריבוי לערך העצמי כי יש לערך עצמי זה כמה וקטורים בלתי תלויים.

נסתכל על וקטור מהצורה: .

ניתן לראות כי : Ly=0y.

אינטואיציה: התת גרף המיוצג על ידי $L_{1,1}$ הוא בעל רכיב קשירות אחד ולכן הוקטור (1,1...,1,1) הוא וקטור עצמי שלו עם ערך עצמי 0.

התרומה של שאר הרכיבים $N_{2,2},\ldots N_{k,k}$ היא 0 כי הגדרנו 0 בשאר הקואורדינאטות. לכן קיבלנו כי (0,...0, 1, ..., 1) הוא וקטור עצמי עם ערך עצמי 0.

גודל רכיב הקשירות הראשון.

באותו האופן, ניתן להגדיר עוד K-1 וקטורים עצמיים נוספים, על ידי שניקח 1 ברכיבים של אותו רכיב קשירות ו 0 בשאר.
בצורה זו נקבל K בחירות בלתי תלויות של הוקטור y.
ולכן קיבלנו בסיס לתת-מרחב המתאים לערך העצמי 0 של מטריצת הלפלסיאן.

<u>משפט 3</u>:במטריצת הסמיכויות הערך העצמי המקסימלי בגרף d רגולרי הוא d.

<u>הוכחה</u>: גרף d רגולרי הוא גרף שבו לכל קודקוד יש d שכנים. כלומר בכל שורה של מטריצת הסמיכויות יש d אחדים. מכיוון שהמטריצה סימטרית גם בכל עמודה יש d אחדים. נניח λ_1 הוא הערך העצמי מקסימלי.

הכיוון $\lambda_1 \geq d$ **:**

נסתכל על וקטור האחדים. $A(1,\ldots,1) = (d,\ldots d) = d(1,\ldots 1)$ מכאן $x = (1,\ldots,1)$ עם ערך עצמי d ולכן הערך עצמי המקסימלי מקיים $\lambda_1 \geq d$.

הכיוון $\lambda_1 \leq d$ **:**

נניח $y = (y_1,\ldots y_n)$ וקטור עצמי ונניח $y_1 \geq y_i, \forall i$. נרצה להראות שהערך העצמי קטן או שווה d. נפעיל על הוקטור y את המטריצה A. $Ay = z$. מכיוון שהמטריצה A היא מטריצת הסמיכויות, משמעות הפעולה היא כי מסכמים את משקלי הקודקודים הסמוכים ל y_1 וזה יהיה הערך של z_1 וכו'.

נניח $Ay = z = (z_1,\ldots z_n)$. כלומר $z_1 = \sum_{i|(1,i)\in E} y_i \leq \sum_{i|(1,i)\in E} y_1 = dy_1$ ולכן $z_1 \leq dy_1$. $\frac{z_1}{y_1} \leq d$

אי שיוויון זה יהיה נכון לכל קואורדינאטה i ולכן הערך העצמי יהיה קטן או שווה ל d.

משני כיוונים אלו נובע כי הערך העצמי המקסימלי הוא בדיוק d.

<u>משפט 4</u>:במטריצת הסמיכויות העי"ע המקסימלי: א. לא יותר גדול מהדרגה המקסימלית.

ב. לא יותר קטן מהדרגה הממוצעת.

<u>הוכחת 4 א'</u>: גרף לא רגולרי הוא גרף שבו לא לכל קודקודים יש את אותה הדרגה.

נרצה להוכיח $\lambda_1 \le d_{\max}$ כאשר λ_1 הוא הערך העצמי המקסימלי.

נניח $y = (y_1,\ldots,y_n)$ הוא וקטור עצמי ונניח $\forall i$, $y_1 \ge y_i$. נרצה להראות שהערך העצמי קטן או שווה ל $d_{\max}$.

נפעיל על הוקטור y את המטריצה A. Ay=z. מכיוון שהמטריצה A היא מטריצת הסמיכויות, משמעות הפעולה היא כי מסכמים את משקלי הקודקודים הסמוכים ל y_1 וזה יהיה z_1 וכו'... נניח, $Ay = z = (z_1,\ldots,z_n)$.

$z_1 = \sum_{i|(1,i)\in E} y_i \le \sum_{i|(1,i)\in E} y_1 \le d_{\max}\cdot y_1$ כלומר $z_1 \le d_{\max}\cdot y_1$ ולכן $\frac{z_1}{x_1} \le d_{\max}$.

אי שוויון זה נכון לכל קואורדינאטה i ולכן הערך העצמי יהיה קטן או שווה $d_{\max}$.

<u>הוכחת 4 ב'</u>:נסמן את הדרגה הממוצעת על ידי $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$ כאשר d_i היא דרגת הקודקודים.

נרצה להוכיח כי $\lambda_1 \ge \bar{d}$ כאשר λ_1 . נוכיח זאת בשני חלקים שמהם תנבע הטענה.

חלק 1': $\lambda_1 \ge y^T Ay$. הטענה נכונה לכל וקטור יחידה $\|y\| = 1$.

חלק 2': נרצה למצוא וקטור יחידה $\|y\| = 1$ שעבורו מתקיים $y^T Ay \ge \bar{d}$.

הוכחת חלק 1': נניח כי $x_1,\ldots,x_n$ הם הוקטורים העצמיים ו $\lambda_1,\ldots\lambda_n$ הם הערכים העצמיים מתאימים להם. $x_1,\ldots,x_n$ מהווים בסיס ל R^n ונניח שכל הוקטורים העצמיים הם מנורמלים כלומר $\left\|x_1^{\,2}\right\| = \left\|x_2^{\,2}\right\| = \ldots = \left\|x_n^{\,2}\right\| = 1$.

נסתכל על הביטוי $y^T Ay$ ונבדוק מהו ערכו עבור וקטור יחידה y . ניתן להציג את y כצירוף לינארי של אברי הבסיס $x_1,\ldots,x_n$: $y = \alpha_1 x_1 + \ldots \alpha_n x_n$. $\|y\| = 1$ ולכן לפי משפט פיתגורס ה-n מימדי $\sum_{i=1}^{n} \alpha^2{}_i = 1$.

$$y^T Ay = \left(\alpha_1 x_1^T + \ldots + \alpha_n x_n^T\right)\cdot A\left(\alpha_1 x_1 + \ldots + \alpha_n x\right) = \left(\alpha_1 x_1^T + \ldots + \alpha_n x_n^T\right)\cdot\left(\alpha_1\lambda_1 x_1 + \ldots \alpha_n\lambda_n x_n\right) = \alpha_1^2\lambda_1 + \ldots \alpha_n^2\lambda_n$$

הסבר: הוקטורים העצמיים ניצבים כי $x_i \cdot x_j = 0$ עבור $i \neq j$.

ולכן קיבלנו כי $\alpha_1^2\lambda_1 + \ldots \alpha_n^2\lambda_n \leq \alpha_1^2\lambda_1 + \ldots \alpha_n^2\lambda_n = \lambda_1(\alpha_1^2 + \ldots + \alpha_n^2) = \lambda_1$. $\lambda_1 \geq y^T Ay$

הוכחת חלק 2': ניקח את y להיות וקטור האחדים המנורמל. כלומר $y = \frac{1}{\sqrt{n}}(1,\ldots,1)$.

A היא מטריצת הסמיכויות. לכן מספר האחדים בה הוא כסכום דרגות הקודקודים.

נחשב את ערך הביטוי $y^T Ay$.

$$y^T Ay = \frac{1}{\sqrt{n}}(1,\ldots,1)\cdot A\frac{1}{\sqrt{n}}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} = \frac{1}{\sqrt{n}}\cdot\frac{1}{\sqrt{n}}(1,\ldots,1)\cdot A\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} = \frac{1}{n}\cdot$$ =סכום דרגות הדוקדוקים $\overline{d}$

הסבר:

$(1,\ldots,1)\cdot A = (d_1,\ldots,d_n)$ כאשר d_i היא דרגת הקודקוד ה i בגרף.

$$\text{סכום דרגת הקודקודים} \quad = \left(d_1,\ldots,d_n\right)\cdot\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} = d_1 + \ldots + d_n$$

הערה: הערך הכי גדול של הביטוי $y^T Ay$ הוא הערך העצמי המקסימלי וזה קורה כאשר נשים במקום y את הוקטור העצמי המתאים לו.

<u>משפט5</u>: בגרף דו-צדדי: גרף שקודקודיו מחולקים לשתי קבוצות זרות U,V כך שכל קשת מחברת קודקוד מ-U עם קודקוד מ-V, הערכים העצמיים באים בזוגות. כלומר אם λ ערך עצמי אז גם $-\lambda$ ערך עצמי.

<u>הוכחה</u>: נניח קודקודי הקבוצה U הם: $p_1,\ldots,p_t$.

וקודקודי הקבוצה V הם: $p_{t+1},\ldots,p_n$.

טענה: אם $(x_1,\ldots,x_t,x_{t+1},\ldots x_n)$ וקטור עצמי עם ערך עצמי λ

אז $(-x_1,\ldots,-x_t,x_{t+1},\ldots x_n)$ וקטור עצמי עם ערך עצמי $-\lambda$.

הוכחה: נראה שהדבר מתקיים עבור נציג p_1 מהקבוצה U, ועבור נציג מהקבוצה V, p_{t+1} .

מההנחה, $\lambda x_1 = \sum_{p_i}^{i} x_i$ כאשר p_i מחובר בקשת ל p_1 .

ולכן גם $-\lambda\cdot(-x_1)=\sum_{p_i}^{i}x_i$ כאשר p_i מחובר בקשת ל p_1 .

כל משקלי הקודקודים מצד ימין של השוויון שייכים לקודקודים מהקבוצה V ולכן משקלם לא משתנה.

בנוסף, $\lambda x_{p+1}=\sum_{p_j}^{j}x_j$ כאשר p_j מחובר בקשת ל p_t .

ולכן גם $-\lambda x_{p+1}=\sum_{p_j}^{j}-x_j$ כאשר p_j מחובר בקשת ל p_t .

כל משקלי הקודקודים מצד ימין של השוויון שייכים לקודקודים מהקבוצה U ולכן משקלם מחליף סימן.

<u>משפט 6</u> : הערכים העצמיים של מטריצת הלפלסאן הם גדולים שווים ל 0.

הלפלסיאן היא מטריצה מוגדרת חיובית (Positive-definite matrix). כלומר סימטרית וללא ערכים עצמיים שליליים.

<u>הוכחה</u> : בגרף רגולרי קל לראות זאת כי :

מטריצת הסמיכויות $Lap=d\cdot I-$

$\Downarrow$

במטריצת הסמיכויות אין ערך עצמי הגדול מ d ולכן במטריצת הלפלסיאן אין ערכים עצמיים שליליים.

בגרף לא רגולרי,

הראינו כי $\lambda_n\le x^TAx\le\lambda_1$ כאשר λ_1 הוא הערך העצמי המקסימלי ו λ_n הוא הערך העצמי המינימלי.

נראה כי לכל וקטור x , $x^TLx\ge 0$ כאשר L היא מטריצת הלפלסיאן. מכאן ינבע כי כל הערכים העצמיים גדולים או שווים ל 0.

ניזכר כי מטריצת הלפלסיאן היא מהצורה :

$\begin{pmatrix} d_1 & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & d_n \end{pmatrix}$ כאשר d_i היא דרגת הקודקוד ה i .

בכל מקום אחר, יש 1- אם ישנה קשת המחברת בין הקודקודים ו 0 אחרת.

$$x^T Lx = (x_1,\dots x_n)\begin{pmatrix} d_1 & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & d_n \end{pmatrix}\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

ניתן לפרק את המטריצה L לסכום של תת מטריצות המייצגות את הקשתות.

$$L = \begin{matrix} & & \text{i} & & \text{j} \\ \text{j} & \begin{pmatrix} 1 & \cdots & -1 \\ \vdots & & \vdots \\ -1 & \cdots & 1 \end{pmatrix} \end{matrix} + \begin{matrix} & & \text{i} & & \text{K} \\ \text{K} & \begin{pmatrix} 1 & \cdots & -1 \\ \vdots & & \vdots \\ -1 & \cdots & 1 \end{pmatrix} \end{matrix} + \dots$$

מייצגת את הקשת (i,k) — מייצגת את הקשת (j,k)

הערה: בשאר הקואורדינאטות יש 0.

נסמן:

$L_{i,j}$ = התת מטריצה המתאימה לקשת המחברת בין i ל j.

$L_{i,k}$ = התת מטריצה המתאימה לקשת המחברת בין i ל k.

ואז $L = \sum L_{n,m}$ כאשר ישנה קשת המחברת בין n ל m.

$$x^T Lx = (x_1,\dots x_n)\cdot\left[L_{i,j} + L_{i,k} + \dots\right]\cdot\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

לפי חוק הפילוג, התוצאה תהיה שווה לסכום ביטויים מהצורה:

$$(x_1,\dots x_n)\cdot L_{i,j}\cdot\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{matrix} & & \text{i} & & \text{j} \\ (x_1,\dots,x_n)\cdot & \text{i} & \begin{pmatrix} 1 & \cdots & -1 \\ \vdots & & \vdots \\ -1 & \cdots & 1 \end{pmatrix} & \cdot & \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \\ & \text{j} & & & \end{matrix} = x_{ii} - x_{ij} - x_{ij} + x_{jj} = (x_i - x_j)^2$$

מכיוון שהסכום מורכב ממחוברים שהם גדולים או שווים ל 0 הוא גם כן גדול או שווה ל 0.

מכאן, למטריצת הלפלסיאן אין ערכים עצמיים שליליים.

<u>משפט 7</u>: AA^T היא מטריצה סימטרית כאשר $A=(a_{ij})$, $A^T=(a_{ji})$.

<u>הוכחה</u>: ניזכר בהגדרת כפל מטריצות:

נניח, $A=(a_{ij})$, $B=\left(b_{ij}\right)$ מטריצות $n\times n$, $C=AB$.

אזי:

$$C_{ik}=\sum_{j=1}^{n}a_{ij}b_{jk}$$

$$C_{ki}=\sum_{j=1}^{n}a_{kj}b_{ji}$$

נרצה שיתקיים: $C_{ik}=C_{ki}$ כאשר: $(a_{ji})=(b_{ij})$. נראה שזה אכן מתקיים.

$$C_{ik}=\sum_{j=1}^{n}a_{ij}b_{jk}=\sum_{j=1}^{n}a_{ij}a_{kj}$$

$$C_{ki}=\sum_{j=1}^{n}a_{kj}b_{ji}=\sum_{j=1}^{n}a_{kj}a_{ij}$$

ומכאן $C_{ik}=C_{ki}$ ולכן $C=AA^T$ היא מטריצה סימטרית כנדרש.

באותה צורה ניתן להראות כי A^TA היא מטריצה סימטרית.

מציאת ערך עצמי מקסימלי של מטריצה בעזרת ה **Power Method**:

שיטה 1: בעזרת ליכסון. מציאת הפולינום האופייני והשורשים שלו. בעזרת שיטה זו מוצאים את כל הערכים העצמיים.

שיטה 2: (The Power Method). אם מעוניינים רק בערך העצמי המקסימלי, ישנה צורה יותר מהירה מהשיטה הקודמת. לא מובטח כי השיטה תעבוד בוודאות, אבל היא תעבוד בהסתברות גבוהה. כלומר זה עובד באופן מעשי. בעזרת שיטה זו מוצאים גם את הוקטור העצמי .
נראה כיצד היא עובדת:

נרצה $(A)(v)=\lambda v$.

ניקח וקטור שרירותי כלשהו r. נפעיל עליו את המטריצה A ונקבל: $Ar=r_1$.

כעת, על מנת להימנע מערכים גדולים בקואורדינאטות, ננרמל את הוקטור r_1.

$$r_1^*=\frac{r_1}{\|r_1\|}$$

נמשיך הלאה ונקבל: $Ar=r_1$

$$Ar_1^*=r_2$$

$$Ar_2^*=r_3 \qquad \text{כאשר} \qquad r_n^*=\frac{r_n}{\|r_n\|}.$$

נניח הוקטורים העצמיים הם: $v_1,v_2,\ldots,v_n$ ונניח שהם מסודרים בסדר עולה. מכיוון שהם מהווים בסיס למרחב הוקטורי R^n,

כל וקטור ניתן להצגה כצרוף לינארי שלהם. בפרט, הוקטור r.

ולכן: $r=\alpha_1 v_1+\alpha_2 v_2+\ldots+\alpha_n v_n$

$$Ar = A(\alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_n v_n) = \alpha_1 A v_1 + \alpha_2 A v_2 + \ldots \alpha_n A v_n$$

$$r_1 = \alpha_1 \lambda_1 v_1 + \alpha_2 \lambda_2 v_2 + \ldots$$

$r_2 = \alpha_1 \lambda_1^2 v_1 + \alpha_2 \lambda_2^2 v_2 + \ldots$ נמשיך כך ונקבל:

$r_k = \alpha_1 \lambda_1^k v_1 + \alpha_2 \lambda_2^k v_2 + \ldots$ מתכנס לוקטור המתאים לערך העצמי הכי גדול אם נחזור על התהליך מספיק פעמים.

אם λ_1 הכי גדול אז כל שאר הרכיבים נהיים זניחים. מכאן מוצאים את הוקטור העצמי.

ככל שחוזרים על התהליך יותר פעמים, מתקרבים יותר לוקטור העצמי.

אחרי שמצאנו את הוקטור העצמי נציב אותו במשוואה ההתחלתית ונמצא את הערך העצמי.
$(A)(v) = \lambda v$.

<u>הערות:</u>

• כאשר בוחרים באקראי את הוקטור r, חשוב שנבחר אותו כך שלא יהיה ניצב לוקטור העצמי המתאים לערך העצמי הגדול ביותר v_1. מכיוון ש $\alpha_i = \langle r, v_i \rangle$, אם r ניצב ל v_1 אז $\alpha_1 = \langle r, v_1 \rangle = 0$. לכן אם המקדם של הערך העצמי הגדול ביותר הוא אפס הוא מתבטל וההתכנסות תהיה לערך העצמי הבא בתור מבחינת גודל. הסיכוי שנעשה בחירה לא מתאימה של הוקטור r הוא מאוד קטן.

• כמה מהר נתכנס לוקטור העצמי? השאלה היא מהם שאר הערכים העצמיים. אם הערכים העצמיים השייכים לאותם וקטור עצמי הם רחוקים, ההתכנסות תהיה מהירה. ואם לא, זה ייקח יותר זמן.

•בעיה: בגרף דו-צדדי. בין הערכים העצמיים של מטריצת הסמיכויות מופיעים בזוגות $d,-d$.

זו אכן בעיה כי אם נעלה בחזקה זוגית, הערך המוחלט שלהם יהיה אותו הדבר.

הרעיון של קליינברג

הרעיון של קליינברג הוא להחשיב כדפים ממליצים רק את הדפים שהם איכותיים בעצמם. אך לשם כך שוב נשאלת השאלה מהו דף ממליץ איכותי?. אחת האפשרויות היא להחשיב כממליץ איכותי דף ששאר הדפים שהוא ממליץ עליהם הם איכותיים גם כן. למעשה נכנסנו לתהליך הבנוי משלבים שחוזרים על עצמם כאשר בכל שלב עלינו לבדוק את איכות דף האינטרנט.

תהליך זה נקרא תהליך איטראטיבי (Iterative). תהליך שבו חוזרים על אותה פעולה שוב ושוב כאשר כל פעולה בודדת נקראת איטראציה. לכן התהליך שקליינברג מציע הוא תהליך איטראטיבי.

קליינברג בעבודתו מתייחס לשני סוגי דפים.

Authority = דף איכותי שנקרא לו דף סמכות. דף בעל איכות טובה עם תוכן מעניין על פי נושא חיפוש מסויים.

Hub =דף רכז. דף שממנו קל להגיע לדפים איכותיים, הוא מהווה ריכוז של הרבה הצבעות כלומר הוא מצביע על הרבה אתרים טובים. לכן הרכזים עוזרים לגלות דפים איכותיים.

לדוגמא, אם אנו מחפשים דפים בעלי איכות טובה עם אינפורמציה על כלבים אז דף רכז יכול להיות פורום של מגדלי כלבים. דף רכז הוא מעין צומת מרכזית שכדאי לעבור דרכה על מנת להגיע למקומות חשובים.

מפאת גודלה של הרשת, לא היינו רוצים להפעיל את אלגוריתם הדרוג על כל הרשת כולה. לכן בדרך כלל יש שלב איתחול הבוחר חלק מהרשת. במאמר לא נאמר בדיוק איך עושים זאת אך בדרך כלל שלב זה מורכב משני חלקים. החלק הראשון הוא סינון הדפים שהכילו את מילת החיפוש בעצמה. החלק השני הוא הרחבת הקבוצה זו על ידי הוספת הדפים שמצביעים עליהם דפים מהחלק הראשון. נרצה לדמות את חלק האינטרנט שבחרנו לגרף דו-צדדי, גרף שקודקודיו מחולקים לשתי קבוצות, דפים איכותיים ודפי רכז. כאשר ישנן קישורים בין שתי קבוצות של דפים אלו אך לא בתוך הקבוצות עצמן. ניתן להסתכל על הדפים כעל קודקודים ועל הקישורים כעל קשתות מכוונות. האינטואיציה לרעיון של קליינברג היא כי אם מופיע באינטרנט מבנה של גרף דו-צדדי צפוף מעין זה, כנראה שמסתתרים כאן דפים בעלי תוכן איכותי. לצורך אישוש ההשערה אפשר פשוט להסתכל מדגמית על כמה מהם ולראות אם אלו דפים איכותיים או לא.

במקרה וחברה לקידום אתרים מגלה את השיטה היא יכולה לבנות מבנה מסוג זה סביב דף אינטרנט שברצונה לקדם. לכל תת גרף כזה נוכל להתאים מטריצה הנקראת מטריצת הסמיכויות. בעזרת הווקטורים העצמיים של מטריצה זו נוכל ליצור דרוג לקבוצת הדפים המיוצגים על ידי הגרף. המאמר מתמודד עם השאלה איך בכלל צריך לגשת לעניין דרוג הדפים. עדיין לא מדובר על איך להפוך את האלגוריתם ליעיל, אלא קודם צריך להבין היטב את השאלה ואת הדרישות מהתשובה בצורה ברורה.

<u>שאילתות ומקורות מוסמכים</u>

שאילתה היא המשפט שהמשתמש כותב בשורת מנוע החיפוש. ישנם שלשה סוגי שאילתות.
סוג א׳: נושא חיפוש ספציפי. זו שאלה ספציפית לדוגמא: מהו מזג האוויר מחר בירושלים?
סוג ב׳: נושא חיפוש רחב. לדוגמא: מצא מידע אודות התכנות בשפת ג׳אווה.
סוג ג׳: בקשה למציאת דפים דומים. לדוגמא: מצא דפים דומים ל java.sun.com. נתמקד בשני הסוגים הראשונים. בכל אחד מהם, הקשיים בחיפוש הם שונים. בחיפוש ספציפי, יש מעט מאוד דפים שיכילו את מילות המפתח המבוקשות ולכן ייתכן שיהיה קשה למצוא את האינפורמציה המבוקשת. לעומת זאת, בחיפוש רחב יהיו אלפים רבים של דפים שיתאימו להגדרת החיפוש. לכן על מנת לאפשר חיפוש אפקטיבי, חייבת להימצא דרך לסנן את הדפים המוסמכים ביותר מתוך המגוון הרב הקיים.
המושג ״דף איכותי״ קשור לשאילתת נושא החיפוש הרחב. מושג זה יהווה את מקור ההתמקדות המרכזי שלנו במאמר. המכשול המרכזי בעניין הוא כיצד נקבע אם דף אינטרנט מסויים הוא דף איכותי עבור שאילתה מסויימת או לא?

<u>ניתוח הרשת המתוקשרת</u>

ניתוח הרשת המתוקשרת יעזור לנו להתמודד עם הקשיים שהעלנו עד כה. למעשה, הקישורים באינטרנט טומנים בתוכם שיפוט אנושי סמוי והטענה שלנו היא שזהו בדיוק השיפוט אותו אנו צריכים כדי להגדיר מהו ״דף איכותי״. לדוגמא, נניח יוצרו של דף p באינטרנט כולל בו קישור לדף q, משמעות הדבר היא כי דף q חשוב לו במידה מסויימת ולכן הוא מהווה עבורו דף איכותי. יותר מכך, הקשרים נותנים לנו את ההזדמנות למצוא דפים איכותיים פוטנציאליים אך ורק על ידי הדפים המצביעים עליהם. העניין מאפשר לנו לעקוף את הבעיה שהועלתה מקודם שכן הרבה דפים אינם מספקים די אינפורמציה על עצמם. ישנם מספר חסרונות בשימוש בקישורים לצורך מטרה זו.
א. קישורים באינטרנט נוצרו מסיבות שונות ולאו דווקא לצורך הקניית איכות לדף.
לדוגמא, ישנם קישורים שמטרתם היחידה היא ניווט באינטרנט וקישורים לצורך פרסומות.
ב. הקושי למצוא את האיזון המתאים בין רלוונטיות ופופולאריות. כאשר שני מושגים אלו תורמים לרעיון האיכות שאנו מדברים עליו. בעבודה זו, נציע מודל המבוסס על קשרים לצורך הקניית המושג דף איכותי ונראה כיצד הוא מוביל לשיטה המזהה דפים רלוונטיים איכותיים

באינטרנט עבור שאילתא כלשהי. המודל מבוסס על הקשר הקיים בין מקורות הסמכות לנושא מסויים לבין אותם דפים המקושרים להרבה מקורות סמכות קרובים. נקרא לדפים מהסוג השני אלו-רכזים. נראה כי סוג מסויים של הקבלה טבעית זו קיימת בין רכזים לדפים איכותיים קיימת בגרף המוגדר על ידי מבנה הרשת המתוקשרת. ננצל זאת על מנת לפתח אלגוריתם שיזהה את הדפים משני הסוגים בו זמנית. האלגוריתם מופעל על תת גרף מרוכז של האינטרנט, אותו נבנה מהפלט שנקבל ממנוע החיפוש. בניית תת גרף זה תייצר אוסף קטן של דפים שיכילו ככל הנראה את הדפים הרלוונטיים ביותר עבור החיפוש.

<u>בניית גרף מרוכז באינטרנט</u>

נוכל להסתכל על אוסף V של דפים מקושרים באינטרנט כעל גרף מכוון. G=(V,E) כך שהקודקודים יהיו הדפים וכל קשת מכוונת $(p,q)\in E$ מראה על קיומו של קישור מ p ל q. נגדיר את דרגת היציאה של קודקוד p להיות מספר הקודקודים שיש לו קישורים אליהם ואת דרגת הכניסה של קודקוד p להיות מספר הקודקודים שיש להם קישורים אליו. מתוך הגרף G נוכל לבודד אזורים קטנים או תתי גרפים באופן הבא: אם $W\subseteq V$ היא תת קבוצה של דפים, נסמן ב$G[W]$ את הגרף המושרה מ W. קודקודיו יהיו הדפים ב W והקשתות שלו יהיו כל הקישורים בין הדפים ב W. נניח כי נתונה שאילתא בעלת נושא חיפוש רחב המצויינת בעזרת מחרוזת σ. נרצה לזהות את הדפים המוסמכים על ידי ניתוח מבנה הקישור. ראשית, אנו חייבים לזהות את תת הגרף באינטרנט עליו יפעל האלגוריתם. מטרתינו היא למקד את המאמץ החישובי בדפים הרלוונטים. לכן, לדוגמא, נוכל להגביל את הניתוח שלנו בקבוצה Q_σ של כל הדפים המכילים את המחרוזת σ. אך יש לכך שני חסרונות חשובים. ראשית, קבוצה זו יכולה להכיל מעל מיליון דפים ולכן תדרוש עלות חישובית רבה. שנית, קודם לכן ציינו כי חלק מהמקורות המוסמכים החשובים ביותר יכולים להיות לא מוכללים בקבוצה זו.

באופן אידיאלי, נרצה להתמקד באוסף S_σ של דפים שהוא בעל התכונות הבאות:

א. S_σ קטנה יחסית.

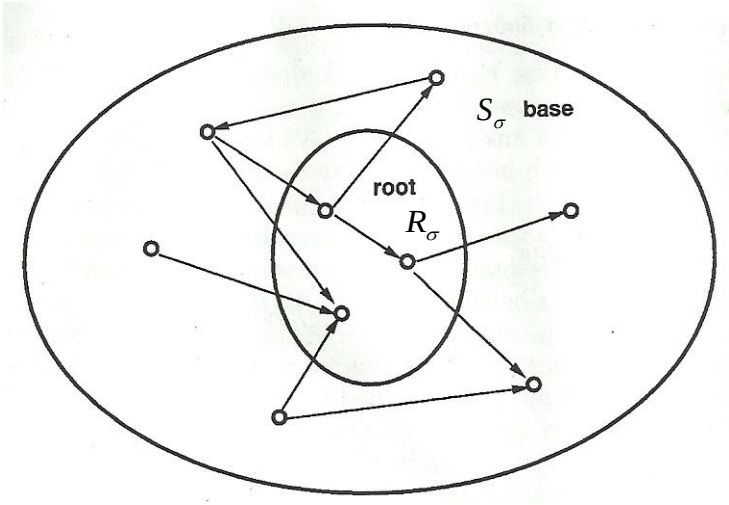
ב. S_σ עשירה בדפים רלוונטיים.

ג. S_σ מכילה את רוב (או הרבה) ממקורות הסמכות החזקים.

על ידי כך שנשמור את S_σ להיות קטנה נוכל להרשות לעצמינו את העלות החישובית על ידי שימוש באלגוריתמים לא טריוויאליים. על ידי הבטחת העושר בדפים הרלוונטיים אנו הופכים את התהליך למציאת מקורות סמכות טובים לקל יותר מכיוון שסביר להניח כי יש אליהם הפנויות רבות בתוך S_σ.

כיצד נוכל למצוא אוסף כזה של דפים? עבור פרמטר t (בדרך כלל הוא בערך 200),קודם לכן ניקח את t הדפים המדורגים ביותר עבור המחרוזת σ על ידי מנוע חיפוש כגון AltaVista או Hotbot.

מנועי חיפוש אלו היו קיימים באותה התקופה, לפני קיומו של Google. נתייחס ל t דפים אלו כאל קבוצה שורשית R_σ. קבוצה זו מקיימת את תנאים א' ו ב' שצויינו קודם אך היא רחוקה מאוד מלקיים את תנאי ג'. על מנת לראות זאת, נשים לב כי t הדפים שבראש הרשימה המוחזרים על ידי מנוע החיפוש כולם יכללו את המחרוזת σ ולכן גם את R_σ. קבוצה זו היא בוודאות תת קבוצה של כל האוסף Q_σ של כל הדפים המכילים את σ. עם זאת טענו כי אפילו Q_σ בדרך כלל אינה מקיימת את תנאי ג'. מעניין לציין כי בדרך כלל יש מעט מאוד קישורים בין דפים ב R_σ, מה שמשאיר אותה למעשה חסרת צורה. לדוגמא, בניסויים שלנו, הקבוצה השורשית עבור השאילתא "java" מורכבת מ 15 קישורים בין דפים מתחומים שונים. נוכל להשתמש בקבוצה השורשית R_σ על מנת לייצר את קבוצת הדפים S_σ שתענה על הדרישות שלנו. אם ניקח מקור סמכות חזק עבור שאילתת חיפוש רחב, ייתכן והוא לא יהיה כלול בקבוצה R_σ, אך סביר להניח שלפחות אחד הדפים ב R_σ יצביע עליו על ידי קישור. לכן נוכל להגדיל את מספר מקורות הסמכות החזקים על ידי הרחבת הקבוצה R_σ. בצורה זו נשיג את הקבוצה S_σ המתקבלת על ידי הרחבת הקבוצה R_σ בעזרת הוספת כל הדפים שישנם קישורים ב R_σ המצביעים עליהם. על מנת להבין זאת נסתכל על הציור הבא:



נרצה לעשות הגבלה על ידי כך שנאפשר לכל דף ב R_σ להביא לכל היותר d דפים המצביעים על דפים שיהפכו ל S_σ. פרט זה הוא חשוב מכיוון שאם נוסיף לקבוצה דפים ללא הגבלה נפגע בתכונה א' הדורשת שהקבוצה S_σ תהיה קטנה יחסית.

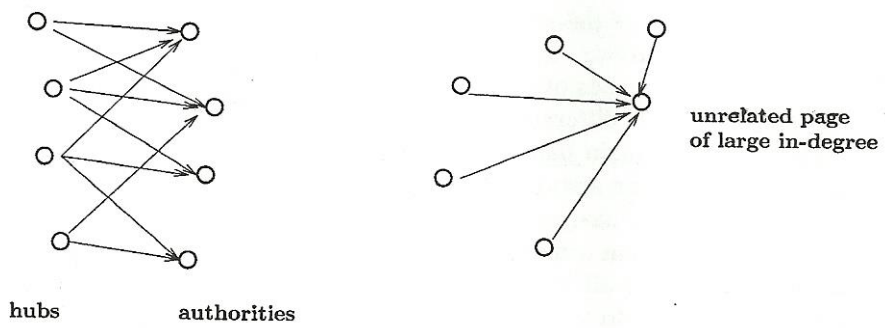
ישנה שיטה שהיא שימושית מאוד לצורך קיזוז השפעת הקישורים שמטרתם היחידה היא ניווט בתוך האינטרנט. נניח ישנו $G[S_\sigma]$ כמו שתיארנו קודם. נבדיל בין שני סוגים של קישורים בתוך $G[S_\sigma]$. קישור חיצוני, הוא קישור המקשר בין דפים עם שמות תחום שונים וקישור פנימי,

המקשר בין דפים עם אותו שם תחום. קישור חיצוני הוא קישור היוצא ומצביע על דף של ארגון אחר. קישור פנימי הוא יוצא ומצביע על דף בתוך הארגון עצמו. הכוונה ב"שם תחום" כאן היא הרמה הראשונה של מחרוזת ה URL הקשורה לעמוד (Domain). כלומר, החלק הראשון של כתובת האינטרנט של הדף. בדרך כלל, מטרתם של הקישורים הפנימיים היא ניווט בתוך התשתית של האתר ולכן הם מספקים הרבה פחות אינפורמציה על הדפים האיכותיים שהם מצביעים אליהם. לכן נמחק את כל הקישורים הפנימיים של הגרף $G[S_\sigma]$ ונשאיר רק את הקשתות המתייחסות לקישורים החיצוניים. התוצאה תהיה הגרף G_σ.

<u>חישוב רכזים ודפים איכותיים</u>

ראינו כיצד ניתן ליצור תת גרף קטן G_σ שהוא מתמקד יחסית בנושא השאילתא. יש בו הרבה דפים רלוונטיים ומקורות סמכות חזקים. עכשיו נראה כיצד לחלץ את מקורות הסמכות מתוך אוסף כל הדפים. נעשה זאת אך ורק על ידי ניתוח הקשרים של G_σ. ניתן לטעון כי השיטה הפשוטה ביותר תהיה לסדר את הדפים על ידי דרגת הכניסה שלהם-מספר הקישורים אשר מצביעים עליהם ב G_σ. קודם לכן, דחינו את הרעיון כי היה מדובר בכל הדפים המכילים את המחרוזת σ. כעת יש לנו אוסף קטן של דפים רלוונטיים שבנינו במפורש, המכיל בתוכו את כל מקורות הסמכות שאנו רוצים למצוא. לכן מקורות סמכות אלו הם גם שייכים ל G_σ וגם יש להם הפניות בצורה רבה על ידי דפים מתוך G_σ. אכן, גישת הדרוג המתבצעת אך ורק על ידי דרוג דרגת הכניסה עובדת טוב יותר בהקשר של G_σ מאשר במסגרת שהצענו קודם לכן. במקרים מסויימים, זה יכול לייצר באופן אחיד תוצאות ברמה גבוהה. בכל זאת, לשיטה עדיין כמה בעיות רציניות. לדוגמא, עבור השאילתא "java" הדפים בעלי דרגת הכניסה הגבוהה ביותר מורכבים מ <u>www.gamelan.com</u> ו java.sun.com יחד עם דפי פרסומות לחופשות באיים הקאריביים ודף הבית של Amazon Books. תערובת זו מציגה סוג של בעיה שיכולה לצוץ בעקבות שיטה של דרוג פשוט. בזמן ששני הדפים הראשונים יכולים להיחשב כתוצאות טובות, השאר אינם רלוונטיות לנושא השאילתא. יש להם דרוג גבוה אך הם חסרות אחדות נושאית. הקושי הבסיסי שנגרם מכך הוא המתח הטמון בגרף G_σ בין מקורות הסמכות החזקים לבין דפים שהם פשוט פופולאריים בצורה אוניברסאלית. אנו מצפים מהסוג השני של הדפים דרגת כניסה גבוהה ללא תלות בנושא השאילתא הבסיסי. ניתן לתהות האם עקיפת בעיות אלו לא תדרוש לעשות שימוש עתידי בתוכן המילולי של של קבוצת הבסיס S_σ יותר מאשר במבנה הקישורי של G_σ. כעת נראה שזה לא המקרה. כלומר למעשה אפשרי לדלות אינפורמציה בצורה אפקטיבית יותר מהקשרים מאשר מהתוכן המילולי. נתחיל מהעובדה הבאה. דפים סמכותיים הרלוונטיים לשאילתא ההתחלתית, לא רק שהם אמורים להיות עם דרגת כניסה גבוהה, אלא מכיוון שהם כולם מקורות סמכות בעלי נושא משותף, אמורה להיות גם חפיפה רבה בקבוצות הדפים המצביעים עליהם. לכן, בנוסף לדפים בעלי הסמכות הגבוהה נצפה למצוא דפים רכזים. דפים רכזים אלו דפים שיש בהם

קישורים לדפים איכותיים רבים. דפים רכזים אלו מקבצים יחד דפים איכותיים בעלי נושא משותף ומאפשרים לנו לזרוק את הדפים הלא רלוונטיים בעלי הדרוג הגבוה. נסתכל על הציור הבא לצורך המחשה.



דפים רכזים ודפים איכותיים מקיימים יחס של חיזוק הדדי. דף רכז טוב הוא דף המצביע על הרבה דפים איכותיים. דף איכותי טוב הוא דף שמצביעים עליו הרבה דפים רכזים טובים. ניתן להגיד בברור כי על מנת לזהות דפים רכזים ודפים איכותיים של תת גרף G_σ אנו צריכים שיטה שתשבור מעגליות זו.

<u>האלגוריתם האיטרטיבי</u>

נשתמש בקשר בין הדפים הרכזים לדפים האיכותיים בעזרת אלגוריתם איטרטיבי המתחזק ומעדכן את המשקלים המספריים של כל דף.

בהינתן גרף מכוון, נחלק משקלים לקודקודים. לדף p נתאים משקל סמכות $x^{\langle p\rangle}$ ומשקל רכזות $y^{\langle p\rangle}$. נרצה שהמשקלים יהיו מנורמלים כלומר $\sum_{p\in S_\sigma}\left(x^{\langle p\rangle}\right)^2=1$, $\sum_{p\in S_\sigma}\left(y^{\langle p\rangle}\right)^2=1$. כעת, נבצע את התהליך האיטראטיבי הבא: תחילה נחלק משקלי רכזות לקודקודים כלומר את $y^{\langle p\rangle}$. למצב זה נקרא y_0. כעת, נבצע לפי התור את הפונקציות הבאות:

$$F: x^{\langle p\rangle} \leftarrow \sum_{q:(q,p)\in E} y^{\langle q\rangle} \quad , \quad Q: y^{\langle p\rangle} \leftarrow \sum_{q:(p,q)\in E} x^{\langle q\rangle} .$$

כלומר, נתחיל מוקטור משקלי רכזות אקראי מנורמל $y_0 = z$ ואז סדר המשקלים יהיה:

$$y_0, x_0, y_1, x_1, y_2, x_2, \ldots$$

למעשה, אם מטריצת הסמיכויות של הגרף היא A אז הפעולות הן:

$$F: x_i \leftarrow A^T y_i \quad , \quad Q: y_i \leftarrow A x_i .$$

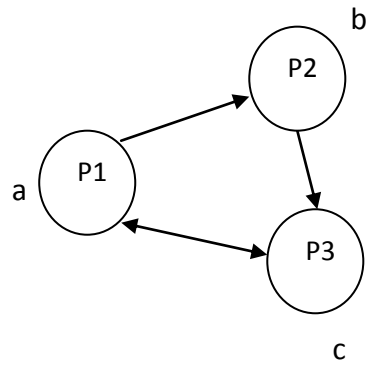
נסביר זאת באמצעות דוגמא:

נסתכל בגרף המכוון הבא, נניח כי הקודקודים הם: p1, p2, p3.

Y = משקל הרכזות. הסכום של אילו שהוא מצביע עליהם.

X = משקל האיכות. הסכום של אילו שמצביעים עליו.

נניח כי משקלי הרכזות הם: (a, b, c).



במקרה זה, מטריצת הסמיכויות היא: $A=\begin{pmatrix}0&1&1\\0&0&1\\1&0&0\end{pmatrix}$ והמטריצה המשחולפת היא

$A^T=\begin{pmatrix}0&0&1\\1&0&0\\1&1&0\end{pmatrix}$. נשים לב כי מטריצת הסמיכויות היא לא סימטרית מכיוון שהגרף הוא מכוון.

טענה: עבור כל מטריצות A^TA ו AA^T הן סימטריות (אך לאו דווקא שוות).

אם נעדכן את משקלי הסמכות נקבל את הוקטור (c, a, a+b) וזה בדיוק:

$A^T\begin{pmatrix}a\\b\\c\end{pmatrix}=\begin{pmatrix}0&0&1\\1&0&0\\1&1&0\end{pmatrix}\begin{pmatrix}a\\b\\c\end{pmatrix}$. כנ"ל אם וקטור משקלי הסמכות היה (a, b, c) והיינו רוצים לעדכן

את וקטור משקלי הרכזות אז באמת היינו מקבלים $A\begin{pmatrix}a\\b\\c\end{pmatrix}=(b+c,c,a)$ כנדרש.

לפי הגדרת הפעולות שדיברנו עליהם: $\begin{matrix}x_i=A^Ty_i\\y_{i+1}=Ax_i\end{matrix}$ כעת,

$$x_{i+1}=A^Ty_{i+1}=(A^TA)x_i$$
.
$$x_{i+1}=(A^TA)x_i$$

מכאן, $x_{i+2}=(A^TA)x_{i+1}=(A^TA)(A^TA)x_i=(A^TA)^2x_i$

$$x_{i+k}=(A^TA)^kx_i$$

$$x_k=(A^TA)^kx_0$$

כמו כן, $x_0=A^Ty_0\equiv A^Tz$

$$x_k=(A^TA)^kA^Tz$$

$y_{i+2}=Ax_{i+1}=AA^Ty_{i+1}$

$$y_{i+2}=AA^Ty_{i+1}$$

$$y_{i+1}=AA^Ty_i$$

$$y_{i+k}=(AA^T)^ky_i$$

$$y_k=(AA^T)^ky_0=(AA^T)^kz$$

הערה: למעשה בכל שלב אנו מנרמלים את הוקטור המתקבל, אם כי לא ציינו זאת במפורש.

נסמן: $AA^T=B$. B סימטרית לפי הטענה. קבלנו כי: $y_k=B^ky_0$.

כל התהליך שעשינו מקביל Power Method רק שהוא מתבצע בשני שלבים. לכן לפי ה Power Method נקבל כי y_k מתכנס לוקטור העצמי y^* של B המתאים לערך העצמי הגדול ביותר של B.

השאלה היא למה זו בדיוק נקודת שבת?

התהליך מורכב משני שלבים:

אם נמצא זוג כך ש: $Ax = \lambda y$ וגם $A^T y = \lambda' x$, נוכל לנרמל על ידי חלוקה ב λ ו λ' בהתאמה

ואז נקבל כמעין נקודת שבת כי: $AA^T y = A \cdot \lambda' x = \lambda' Ax = \lambda' \lambda y$ כלומר $By = \lambda' \lambda y$ ו

$\frac{1}{\lambda' \lambda} By = y$. לכן y היא נקודת השבת של המטריצה $\frac{1}{\lambda' \lambda} B$.

המסקנה היא כי כל התהליך מקביל ל Power Method.

כמו כן ניתן להגדיר מטריצה $C = A^T A$ ובאותה הדרך להגיע למצב $x_k = C^K x_0$ ואז

$A^T Ax = \lambda \lambda' x$ כלומר $Cx = \lambda \lambda' x$, $\frac{1}{\lambda \lambda'} Cx = x$.

הגענו לתוצאה כי ל B ו C אותם ערכים עצמיים אך וקטורים עצמיים שונים.

מסקנה: בעזרת תהליך ה Power Method הגענו לוקטור המשקלים המתאים לערך העצמי המקסימלי. וקטור זה מהווה נקודת שבת ולכן נקבל דרוג לקודקודים.

הערה: האלגוריתם שהוצג כאן הוא רק אחד מתוך כמה שיקולים בעזרתם מנועי החיפוש מבצעים את הדרוג. הדרוג הסופי נבנה בעזרת איזון בין כמה שיקולים.
לפי התוצאות המוצגות במאמר ההתכנסות של Iterate היא די מהירה. למעשה, הממצאים הם כי מספיקות כ-20 אינטראציות כדי להתכנס להגיע למצב שבו כמעט ולא זזים מנקודת השבת. אחרי שהתכנסנו לנקודת השבת ניתן להגדיר את המספר c . מספר זה ייצג את כמות הדפים הראשונים בדרוג שנרצה שיוצגו על ידי מנוע החיפוש.

רשימה ביבליוגרפית

•JON M. KEINBERG/Authoritative Sources in a Hyperlinked Environment/*Journal of the ACM, Vol. 46, No. 5, September 1999, pp. 604-632.*

• FAN R. K. CHUNG/ Spectral Graph Theory, 1994.