אלגוריתמי קירוב מקומי לדירוג דפים (PageRank) ודירוג דפים הפוך

לי-טל משיח

אלגוריתמי קירוב מקומי לדירוג דפים
(PageRank) ודירוג דפים הפוך

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים במדעי המחשב

לי-טל משיח

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

שבט תשס"ח חיפה פברואר 2008

המחקר נעשה בהנחיית זיו בר-יוסף בפקולטה למדעי המחשב

אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי

ברצוני להביע את תודתי הכנה למנחה שלי, ד"ר זיו בר-יוסף. אני מודה לזיו על כך שחשף אותי לעולם מרתק, עולם החיפוש. אני מודה לו על אווירת עבודה נפלאה, על כך שתמיד עזר לי לשמור על גישה חיובית על אף הקשיים ועל ההנחייה הצמודה שזכיתי לקבל.

תודה רבה להוריי האהובים, יולנדה ואלי משיח, על תמיכתם האין-סופית, על העידוד ועל כך שתמיד האמינו בי. ברצוני גם להודות לבן-זוגי, גבי קליאוט, על כך שהוא עומד לצידי בעליות ובמורדות החיים. עבודה זו מוקדשת להם.

מבוא

דירוג דפים (פייג'ראנק) הינו אלגוריתם ידוע לדירוג דפי היפרטקסט המקושרים ביניהם, המשמש מנועי חיפוש, בינהם גוגל, לדירוג תוצאות חיפוש. פייג'ראנק, אשר הוצג לראשונה ע"י פייג' ושות' [5], מבוסס על ההגדרה הרקורסיבית, שדף הינו "חשוב" אם דפים "חשובים" מצביעים אליו. במהלך העשור האחרון פורסמו מאמרים רבים הדנים בדרכים לחישוב יעיל של פייג'ראנק.

אלגוריתמי קירוב מקומי לפייג'ראנק. רוב האלגוריתמים לחישוב פייג'ראנק, בין אם הם ריכוזיים, מקביליים או מבוזרים, התמקדו בחישוב גלובאלי של וקטור הפייג'ראנק, כלומר בחישוב דירוג הפייג'ראנק לכל הצמתים בגרף. אך בעוד שלאפליקציות רבות חישוב פייג'ראנק גלובלי באמת הכרחי, פעמים רבות נתעניין בדירוג פייג'ראנק לקבוצה קטנה של צמתים בלבד.

לדוגמה, נתבונן בחברה אשר בבעלותה אתר אינטרנט והיא מעוניינת לקדם את האתר בדירוג מנועי החיפוש מתוך כוונה למשוך תנועת לקוחות פוטנציאלים. ידוע כי פייג'ראנק הוא מדד חשוב בפונקצית הדירוג של מנועי החיפוש העיקריים. לכן, מעקב אחר הפייג'ראנק של האתר יאפשר למנהל האתר הבנה עמוקה של מצב האתר בדירוג מנועי החיפוש ולאחריה אפשרות לנקיטת צעדים לשיפור הפייג'ראנק. במיקרה כזה, מנהל האתר יתעניין בדירוג הפייג'ראנק של האתר שלו בלבד (ואולי בנוסף בדירוגי אתרים מתחרים), אבל לא בדירוג הפייג'ראנק של כל שאר הדפים ברשת האינטרנט.

רוב מנועי החיפוש בוחרים להשאיר את מדד הפייג'ראנק כדבר חסוי. חלק ממנועי החיפוש מפרסמים ערכי פייג'ראנק גסים (לדוגמה, בסרגל הכלים של Google), אך אלו ניתנים בדרך כלל בסולם לוגריתמי מ-1 עד 10. משתמשים המעוניינים להשיג דירוג פייג'ראנק מדוייק יותר לדפים לפי בחירתם, נותרים ללא ברירה אלא לחשב זאת בעצמם. לרב המשתמשים, חישוב פייג'ראנק גלובלי אינו בא בחשבון כיוון שהדבר דורש משאבים וידע מקצועי רב. הדבר מעלה את השאלה הטבעית הבאה: האם ניתן לחשב את דירוג הפייג'ראנק של דף אינטרנט בודד בכמות משאבים סבירה?

Chen, Gan ו-Suel [1] הציגו לראשונה את בעית הקירוב המקומי של פייג'ראנק. נניח כי ניתנת לנו גישה לגרף גדול G (לדוגמא, גרף האינטרנט) דרך שרת אשר לכל שאילתה על צומת x מחזיר את הקשתות היוצאות והניכנסות ל-x (להלן שרת לינקים). במיקרה שבו G הוא גרף האינטרנט, ניתן לחלץ את הקשתות היוצאות מצומת x פשוט מתוך דף x וניתן לקבל את הקשתות הניכנסות ע"י שליחת שאילתת link: למנוע חיפוש.

האם נוכל כעת, בעזרת מספר מצומצם של שאילתות למנוע החיפוש, להעריך בדיוק גבוה את הפייג'ראנק של צומת מטרה x?

Chen ושות' הציעו אלגוריתם לפתרון בעייה זו. האלגוריתם מבצע crawl לתת-גרף קטן בסביבת צומת המטרה, מיישם מספר היוריסטיקות לניחוש מדד הפייג'ראנק של צמתים על שפת תת-הגרף ולבסוף מחשב את הפייג'ראנק של צומת המטרה. Chen ושות' הראו בצורה אמפירית כי אלגוריתם זה מספק קירוב טוב בממוצע. אך למרות זאת, הם ציינו במפורש, כי צמתים עם דרגת כניסה גבוהה גורמים לעיתים לאלגוריתם להיות מאוד מאוד לא מדוייק. I

חסמים תחתונים. בעבודה זאת חקרנו את מגבלות הקירוב המקומי של פייג'ראנק. זיהינו שתי סיבות לכך שהקירוב המקומי יהיה קשה לחישוב: (1) קיומן של צמתים עם דרגת כניסה גבוהה ו- (2) התכנסות איטית של ההילוך המיקרי של פייג'ראנק.

כדי להדגים את ההשפעה של דרגת כניסה גבוהה, עבור כל n, הצגנו משפחה של גרפים בגודל n עם דרגת כניסה מקסימלית גבוהה ($\Omega(n)$) ואשר עבורם כל אלגוריתם ידרש לשלוח $\Omega(\sqrt{n})$ שאילתות לשרת על מנת להשיג קירוב פייג'ראנק מדוייק. עבור n-ים גדולים, משיכת $\sqrt{n}$ דפים מהרשת או שליחת $\sqrt{n}$ שאילתות למנוע חיפוש הינן פעולות יקרות ביותר (לדוגמה, עבור גרף האינטרנט $n \geq 10B$, ומכך $\sqrt{n} \geq 128K$). החסם התחתון אותו הוכחנו מתייחס גם לאלגורתימים הסתברותיים וגם לאלגוריתמים דטרמיניסטיים. עבור אלגוריתמים דטרמיניסטיים אנו מוכיחים חסם תחתון חזק יותר (ואופטימלי) $\Omega(n)$.

באופן דומה, הדגמנו את השפעת התכנסות איטית של פייג'ראנק ע"י הצגת משפחה של גרפים עליהם התכנסות מיקרית של פייג'ראנק הינה איטית ביותר (ב- $\Omega(\log n)$ צעדים) ועליהם כל אלגוריתם דורש שליחת $\Omega(n^{\frac{1}{2}-\varepsilon})$ שאילתות, במטרה לקבל קירוב טוב לפייג'ראנק ($\varepsilon > 0$ הוא קבוע קטן אשר תלוי בקבוע הדעיכה של פייג'ראנק). שוב, חסם תחתון זה רלוונטי לאלגוריתמים הסתברותיים ודטרמינסטיים. עבור אלגוריתמים דטרמיניסטיים הראנו חסם תחתון אופטימלי של $\Omega(n)$.

שני החסמים התחתונים הללו הינם בלתי תלויים: משפחת הגרפים הקשים שניבנתה לחסם התחתון הראשון הינה בעלת התכנסות מהירה של פייג'ראנק (שתי איטרציות) בעוד שהמשפחה השנייה של הגרפים אשר ניבנתה לחסם התחתון השני הינה בעלת דרגת כניסה חסומה (לכל היותר 2).

חסם עליון. לאחר שהוכחנו כי קירוב מקומי של פייג'ראנק הינו קשה עבור גרפים בעלי דרגת כניסה גבוהה או כאלו שפייג'ראנק אינו מתכנס עליהם במהירות, היה זה אך טבעי לשאול האם קירוב מקומי של פייג'ראנק אפשרי על גרפים עם דרגת כניסה חסומה והתכנסות מהירה של פייג'ראנק. הראנו כי האלגוריתם של Chen ושות' עובד היטב, על גרפים שכאלו. הוכחנו שאם ההילוך המיקרי של פייג'ראנק מתכנס על הגרף ב-r צעדים ודרגת הכניסה המקסימלית של הגרף הינה d, אז האלגוריתם של Chen ושות' דורש לכל היותר d^r שאילתות לשרת הלינקים.

II **נק לעומת פייג'ראנק הפוך.** היות וגרף האינטרנט עמוס בצמתים עם דרגת כניסה גבוהה, החסם התחתון הראשון שלנו מראה כי לעיתים קרובות קירוב מקומי של פייג'ראנק יהיה בלתי ניתן לביצוע על גרף האינטרנט. אימתנו אבחנה זו ע"י אנליזה אמפירית של 280,000 צמתים להם ביצענו crawl מתוך אתר www.stanford.edu. הראנו כי קירוב מקומי של פייג'ראנק הינו קשה במיוחד לצמתים בעלי פייג'ראנק גבוה. צמתים אלו דורשים אלפי שאילתות לשרת הלינקים. ממצאים אלו מספקים הסבר אנליטי ואמפירי לקשיים ש-Chen ושות' נתקלו בעבודתם.

לאחר מכן הראנו שגרף האינטרנט ההפוך (הגרף המתקבל ע"י הפיכת כיווני הקשתות) מתאים הרבה יותר לקירוב מקומי של פייג'ראנק. ע"י ניתוח ה-crawl של www.stanford.edu, הראנו שגרף האינטרנט ההפוך, בדומה לגרף האינטרנט, מקיים התכנסות מהירה של פייג'ראנק (עבור 80% מהצמתים בגרף ההפוך, פייג'ראנק מתכנס בלכל היותר 20 איטרציות). בנוסף הראנו כי לגרף ההפוך דרגת כניסה חסומה (דרגת הכניסה המקסימלית הינה 255 בלבד לעומת 38,606 בגרף הרגיל). אנליזה עדינה יותר מראה כי קצב הגידול של crawl בסביבת צמתים עם פייג'ראנק גבוה הינו איטי בהרבה על הגרף ההפוך מאשר על הגרף הרגיל (ב-80% בממוצע). ממצאים אלו מרמזים על כך שהאלגוריתם של Chen ושות' עשוי לעבוד טוב יותר על גרף האינטרנט ההפוך.

על מנת שנוכל להעמיד היפותזה זאת למבחן, מדדנו את ביצועי האלגוריתם של Chen ושות' על דגימות של צמתים מתוך הגרף של www.stanford.edu. הראנו שעבור צמתים המדורגים גבוה הביצועים של האלגוריתם על הגרף ההפוך הם עד פי שלושה טובים יותר מאשר על הגרף הרגיל.

המסקנה מכך היא שהגרף ההפוך מתאים יותר לקירובים מקומיים של פייג'ראנק מאשר הגרף הרגיל. לכן, חישוב פייג'ראנק הפוך (פייג'ראנק על הגרף ההפוך) ישים יותר מאשר חישוב מקומי של פייג'ראנק רגיל.

יישומים של פייג'ראנק הפוך. אומנם קירוב מקומי של פייג'ראנק הפוך הינו קל יותר מאשר קירוב מקומי של פייג'ראנק רגיל, אבל למה לנו לחשב זאת מלחתחילה?

אנו מבחינים שלפייג'ראנק הפוך קיימות מספר רב של אפליקציות בגרף האינטרנט ובגרפים אחרים עם תכונות דומות. בעבר השתמשו כבר בפייג'ראנק הפוך לבחירת seeds טובים למדד ה-TrustRank [3], למציאת צמתים משפיעות ברשתות חברתיות [4] ולמציאת hubs בגרף האינטרנט [2]. אנו מציגים שני יישומים מקוריים חדשים לפייג'ראנק הפוך: (1) מציאת seeds טובים ל-crawling ו-(2) מדידת קירבה סמנטית בין שני מושגים שונים בטקסונומיה.

ביבליוגרפיה:

[1] Y. Chen, Q, Gan, and T. Suel. Local methods for estimating pagerank values. In Proc. CIKM, pages 381-389, 2004.

[2] D. Fogaras. Where to Start Browsing the Web? In IICS, pages 65-79, 2003.

[3] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating Web Spam with TrustRank. In VLDB, pages 576-587, 2004.

[4] A. Java, P. Kolari, T. Finin, and T. Oates. Modeling the Spread of Influence on the Blogosphere. Technical report, University of Maryland, Baltimore Country, 2006.

[5] L. Page, S. Brin, R. motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.