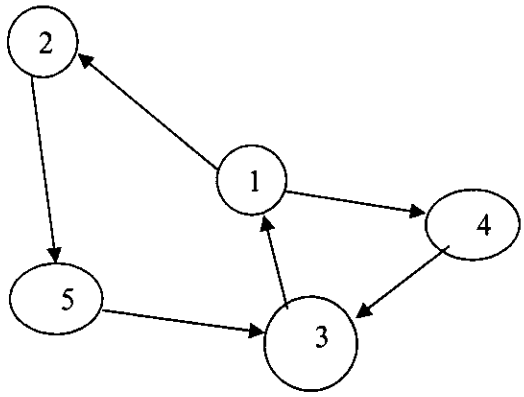
אוניברסיטת בן-גוריון - המחלקה להנדסת מערכות מידע
קורס איחזור מידע וספריות דיגיטליות
סמסטר אביב תשס"ו- 09.07.06 – מועד א

מרצה: ד"ר ברכה שפירא, אסיסטנט : ארז שלום
משך המבחן: 3 שעות , חומר עזר מודפס או כתוב - מותר, מחשבון- מותר
יש לענות על כל 5 השאלות. **יש להחזיר את השאלונים.**

1. (15%) נתונה רשת (בציור) המתארת צמתי Web והצבעות ביניהם :

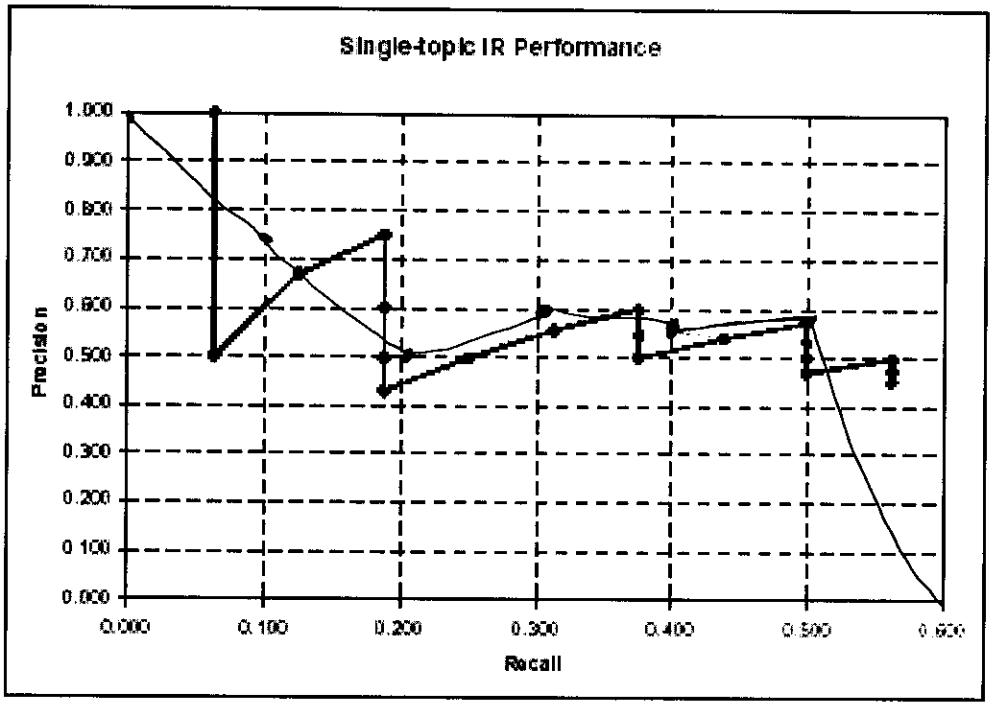


א. (3%) אם נוסיף קישור מצומת 2 לצומת 4, האם וכיצד יושפעו ערכי HUB ו – AUTHORITY (על פי אלגוריתם HITS) של צומת 5 (כלומר , האם הערכים יהיו גבוהים, או נמוכים יותר, או ללא שינוי ומדוע) .

ב. (7%) חשב שתי איטרציות של pagerank של הגרף שבציור לאחר ההוספה מסעיף א. (ללא נרמול).
הערך איזה צומת תהיה בעלת ערך pagerank גבוה ביותר לאחר ההתכנסות. הסבר את הערכתך.

ג. (5%) האם אלגוריתמים של link-analysis כדוגמת pagerank מודדים גם את פופולאריות הדף אצל הגולשים. אם כן, כיצד? ואם לא איך אפשר למדוד פופולאריות של דפים אצל גולשים?

2.(30%) הגרף הבא מתאר תוצאה של שאילתא אחת במנוע חיפוש שבה הוחזרו 20 תוצאות. ידוע שלשאילתא יש 16 תוצאות רלוונטיות (שלא כולן הוחזרו על ידי המנוע). הנקודות על הגרף מראות את ה-precision ו –ה -recall בכל נקודות המסמכים שהוחזרו. (כלומר כל נקודה מייצגת precision ו-recall למסמכים שהוחזרו עד נקודה זו על ידי המנוע).



תשובה
ל ג-ד

א. (6%) יש לצייר על גבי הגרף הקיים את הגרף בנקודות ה recall הסטנדרטיות באמצעות אינטרפולציה (בשיטת האינטרפולציה שנלמדה בקורס)

ב. (6%) אילו מסמכים מבין המסמכים שהמנוע החזיר היו רלוונטים? – התשובה צריכה לכלול מספרי מסמכים לפי סדר דירוגם על ידי המנוע. (למשל אם התשובה היא- 1,3,5 - הכוונה היא שהמסמכים שדורגו במקומות האלו על ידי המנוע הם רלוונטים).

ג. (6%) חשב/י את ה precision הממוצע (Mean Average Precision) (הראה את דרך החישוב)

ד. (3%) חשב/י את R-Precision

ה. (3%) מהו ה Precision בנקודת 10 מסמכים?

ו. (6%) תאר שתי חולשות של מאגרי הבדיקה המסופקים על ידי TREC.

(20%) בקובץ Posting של אינדקס הופכי מסויים כל כניסה כוללת: [docid, frequency of term]
נתון האינדקס של מאגר במבנה שהוזכר למעלה ובו 5 מסמכים:

Dictionary	Posting
Mother--→	{1,3},{3,2},{4,1},{5,3}
Father-- →	{2,4}, {3,1},{5,2}
Family---→	{2,1}, {4,3}
Brother--→	{1,4},{3,3},{4,2}

א. (4%) אילו מסמכים יוחזרו ובאיזה סדר לשאילתות הבאות, במנוע חיפוש במודל בוליאני טהור:
1) Mother and Family
2) Mother or Family

(1) א. קשר מדומה 3 לדומה 4 יעלה את ה-HUB של 3 ובמקביל יפגע את ה-authority של 5.
ה-HUB של 5 "יגיע אל 3 שינוי.

ב. איטרציה 1.

דומה 1	דומה 2	דומה 3	דומה 4	דומה 5
$0.15+0.85=1$	$0.15+0.85/2=$ 0.575	$0.15+0.85+0.85=$ 1.85	$0.15+0.85/2+0.85/2=1$	$0.15+0.85/2=$ 0.575
$0.15+1.85*0.85=$ 1.7225	$0.15+0.85/2=$ 0.575	$0.85+0.85*0.575+0.15=1.48875$	$0.85/2+0.85/2+0.575+0.15=0.819$	ארץ לדומה 3: $\frac{0.85}{2}*0.575+0.15=$ 0.394

אפשרויות נבחרות אחרות לחישוב איטרציה עם ערך של 0.2 לכל דומה.

3 במקום 0.15 $\frac{\text{צוים}}{3}$ $\frac{(1-d)}{m}$

הדומה בעלת ערך pagerank גבוה היא דומה מס' 3 או 1.
(שתי התשובות התקבלו בנבחנות אבל שדומה 3 נשאלת בבחינה יותר).
הנימוק: השאלה של דומה 3 ל 4,5 - דומה 3 מקבלת מישהו ולכן
בבחינה מכל אחת מהן.
השאלה של דומה 3 ל 3 - דומה 3 מקבלת רק 1/2 מדומה 1
ואילו דומה 3 מקבלת את השאלה של דומה 1 השלמה (דרך דומה 4,5).

דומה 3 מצביעה לדומה 1 ויש לה את ה RK. לכן לאחר ההתכנסות
הם בערך באותו גובה.

1

ג) אלגוריתם של Kaczynski-Agrawal לא מודדים באלפן ושיר את הפופולריות של הדף אלא אלשים. אלגו' את התשובות שלו כן הדפים בשלמות. אם כי באלפן קיים אפשר לומר שדפים פופולריים סביר שיקבלו מלא דפים אחרים באלפן גבוה.

בכ' למדוד באלפן ישר פופולריות של דפים יש לבדוק מספר כניסות אל דפים.

2

ב'

מסמכים רלוונטיים (אפשר למצוא מספר מלא של הדרך).

1, 3, 4, 8, 9, 10, 13, 14, 18

ג' נקודות Recall סטנדרטי:

נקודת Recall	precision	
0	1/1	=1
0.1	3/4	=0.75
0.2	4/8	0.5
0.3	6/10	0.3
0.4	7/13	0.538
0.5	8/14	0.57
0.6	0	
⋮	⋮	
1	0	

מחלקים מספר 11 הנקודות הנ"ל.

$$\frac{1+0.75+0.5+0.3+0.538+0.57}{11} = 0.3599$$

(2)

ג. R-precision = $8/16 = 0.5$

ד. precision@10 = $\frac{6}{10}$

ו. תלמידים לא יכולים [redacted] שאלות לא אחיד.

ב. הדגישה ה recall

3 א. ניתן לקבל מסמך 4 מכיוון שהוא המסמך היחיד שמכיל את שני המילים. (המילה בלתי נדירה)

ב. נחשב כל המסמכים, כולל סדר המילים.

ג.

$$W_{mother,1} = \frac{3}{4} \cdot \frac{5}{4} = \frac{15}{16}$$

(f_{ij} / $\max[f_{ij}]$, N / df_i)

$$W_{mother,3} = \frac{2}{3} \cdot \frac{5}{4} = \frac{10}{12}$$

$$W_{mother,4} = \frac{1}{3} \cdot \frac{5}{4} = \frac{5}{12}$$

$$W_{m,5} = \frac{3}{3} \cdot \frac{5}{4} = \frac{5}{4}$$

$$W_{father,2} = \frac{4}{4} \cdot \frac{5}{3} = \frac{5}{3}$$

$$W_{F,3} = \frac{1}{3} \cdot \frac{5}{3} = \frac{5}{9}$$

$$W_{F,5} = \frac{2}{3} \cdot \frac{5}{3} = \frac{10}{9}$$

$$Sim(D_i, q) = \text{Inner product} = \sum w_{ij} \cdot q_{ij}$$

$$Sim(D_1, q) = \frac{15}{16} = 0.9375$$

$$Sim(D_2, q) = \frac{5}{3} = 1.66$$

$$Sim(D_3, q) = \frac{10}{12} + \frac{5}{9} = 1.38$$

$$Sim(D_4, q) = \frac{5}{12} = 0.416$$

$$Sim(D_5, q) = \frac{5}{4} + \frac{10}{9} = 2.36$$

המסמכים יוחזרו בסדר הבא:

$$D_5 \rightarrow D_2 \rightarrow D_3 \rightarrow D_1 \rightarrow D_4$$

(3)

(ג) התוצאות לא ישתנו, כיוון שגם בפעם הבאה בפונקציה יכן שבוצע שינוי, דבר שישנה את מספרי הדירוג, אין לא ישפיע על אופן הדירוג.

(3)

(3) למעשה 2 דברים יקרו: (1) ישנה ה idf ב אחוז $(\frac{5}{3})$

(2) ה tf ב עמוד 5 ישתנה.

לכן ניתן לראות שם מתשמים שגם 2, 3, 5 יתחלו הראשונים

לא כמו קודם, ובגלל השינוי. (כאשר 5 יתן עם 2 ו 3 יתחלו הראשונים)

(4) פונ מרחק: $f(\vec{x},\vec{y}) = |x_1-y_1| + |x_2-y_2|$ $k=2$

cluster 1	cluster 2
d_1	d_2-d_7

$f(d_1,d_3) = |0.8-0.4| + |0.4-0.2| = 1.1$

$f(d_2,d_3) = 0.9$

$f(d_2,d_4) = 1$

$f(d_1,d_5) = 1$

$f(d_2,d_5) = 0.8$

$f(d_1,d_6) = 1$

$f(d_2,d_6) = 0.8$

$f(d_1,d_7) = 0.8$

$f(d_2,d_7) = 0.6$

$C_1 = (0.9, 0.8)$

$C_2 = \frac{1}{6}(d_2+d_3+d_4+d_5+d_6+d_7) = \frac{1}{6}(0.51 \quad 0.31)$

(4)

cluster 1	cluster 2
d_1	d_3
d_2	d_4
	d_5
	d_6
	d_7

$f(d_1, c_2) = 0.88$

$f(d_1, c_1) = 0$

$f(d_2, c_1) = 0.2$

$f(d_2, c_2) = 0.68$

$f(d_3, c_1) = 1.1$

$f(d_3, c_2) = 0.4$

$f(d_4, c_1) = 1.1$

$f(d_4, c_2 = 0.4$

$f(d_5, c_1) = 1$

$f(d_5, c_2 = 0.12$

$f(d_6, c_1) = 1$

$f(d_6, c_2) = 0.12$

$f(d_7, c_1) = 0.8$

$f(d_7, c_2) = 0.1$

$C_1 = \frac{1}{2}(d_1 + d_2) = (0.75 \quad 0.75)$

$C_2 = \frac{1}{5}(d_3 + d_4 + d_5 + d_6 + d_7) = (0.46 \quad 0.24)$

cluster 1	cluster 2
d_1	d_3
d_2	d_4
	d_5
	d_6
	d_7

$f(d_1, c_1) = 0.1$

$f(d_1, c_2) = 1$

$f(d_2, c_1) = 0.1$

$f(d_2, c_2) = 0.8$

$f(d_3, c_2) = 0.42$

$f(d_4, c_1) = 1$

$f(d_4, c_2) = 0.38$

$f(d_5, c_1) = 0.9$

$f(d_5, c_2) = 0.08$

$f(d_6, c_1) = 0.9$

$f(d_6, c_2) = 0.2$

$f(d_7, c_1) = 0.7$

$f(d_7, c_2) = 0.2$

לא איכר כלום

⇓

הסדרתי

ב) הבעיות הן: מיון לא-מאוזן, תיוג אין יותר מדגם (אותו תיקון תשובה)

ג) 1) clustering - כלי יעזור מחשוב את הקטגוריות המנסה לקבץ
לפי אלגוריתם את הנתונים.

2) categorizing - יעזור מחשוב את הקטגוריות לפי מסווגים
את הנתונים.

3) training-set יכול יעזור, משהו שם clustering אין משמעות
מתרשמת מלא נכונות" ש training set ומה לעזור מהם.

4) feature selection - יכול לסייע, לשנות המקום לקחת טבע פרמטרים
ניתן לקחת כן כ ובכן לשפר בכוחם.

5) הוספת מודול שמחולל את תוכנו לפי נושאים. בנוסף כאשר קוראת ה לאתר ניתן לבקש את ה domain של הלקוח.
ניתן להכניס את המודול בין ה index ל search מודול כזה.
(או לקחי בין ה לאתר כמומחה)

ב) א. mirror sites
ב. security של הנתונים.
ג. דפים דינמיים וכו'.

ג) אלגוריתם אופטימלי לא הצמן את התשובות המוחלטות לא.
לא ניתן התשובות האלגוריתם מבוסס את המצב הסטטיסטי
של המשתמש שיבקר להחליט מה רב של דפים.
בעזרת אלגוריתם סטטיסטי מסתמן על ניתוח סטטיסטי ע"י תגובות המנוים
בעבר אפשר בכל לא מאגרים כלשהם וכן המנוים התגובות בו
בעתיד אין בהכרח קשורות לנושא.