מודלים במסחר אלקטרוני - תרגיל בית 2

תאריך הגשה: 20/6/2007

המרצה: פרופ' משה טננהולץ. המתרגל: מר אלון אלטמן.

שאלה 1

שאלה זו עוסקת ב- market-basket.

נתונים הסלים הבאים ומספר הלקוחות שקנו כל סל:

סל	מספר לקוחות	סל	מספר לקוחות
$\{C\}$	1,000	$\{A,B,C\}$	1,100
$\{D\}$	2,000	$\{A,D\}$	500
$\{E\}$	500	$\{B,E\}$	2,000
$\{A,B\}$	3,000	$\{B,D,E\}$	1,800
$\{A,C\}$	1,200	$\{A,C,D,E\}$	100
$\{B,C\}$	500		

שימו לב: ספירת הלקוחות של תת-סל לא כוללת את הלקוחות שקנו סלים שמכילים אותו. סלים שאינם מצויינים לעיל לא נקנו כלל.

מעוניינים להשתמש באלגוריתם PCY למציאת frequent pairs, כאשר זוג נפוץ הוא כזה שקנו אותו לפחות 4,000 לקוחות.

א. אילו מוצרים עומדים בתנאי A priori? יש לפרט את החישוב.

ב. נתונה פונקציית ערבול h המערבלת את הזוגות לדליים לפי הטבלה הבאה:

דלי 1	דלי 2	דלי 3	דלי 4	דלי 5
$\{A,B\}$	$\{A,C\}$	$\{A,D\}$	$\{A,E\}$	$\{C,D\}$
$\{C,E\}$	$\{B,D\}$	$\{B,E\}$	$\{B,C\}$	$\{D,E\}$

בשלב השני של האלגוריתם משתמשים בפונקציית הערבול h על הזוגות שנותרו מהפעלת A priori. אילו דליים ייחשבו נפוצים? יש לפרט את החישוב.

ג. בשלב השלישי סופרים את הזוגות מהסלים הנפוצים. מהם הזוגות הנפוצים בסיום האלגוריתם?

שאלה 2

שאלה זו עוסקת באלגוריתם multistage PCY למציאת frequent pairs.

א. הוצע, לשם יעילות, להשתמש באותה פונקציית hash בכל שלבי האלגוריתם. כיצד תשפיע פעולה זו על ביצועי האלגוריתם? נמקו את תשובתכם!

ב. לשם חסכון בזיכרון, הוצע שבכל שלב יישמר רק ה- bitmap מהשלב הקודם ולא יישמרו bitmaps אחרים ו/או רשימת ה- frequent items. כיצד תשפיע פעולה זו על ביצועי האלגוריתם? נמקו את תשובתכם!

ג. הוצע לשפר את האלגוריתם ולשמור בכל שלב את תוצאות הספירה עצמה של ה- hash table ולא רק bitmap. כיצד ישפיע שינוי זה על ביצועי האלגוריתם, אם אנו מניחים שכל הזיכרון הפנוי מנוצל לאחסון ה- hash table?

שאלה 3

שאלה זו עוסקת ב- Min Hash.

נתונה טבלת רכישות מוצרים על ידי לקוחות ונתונות 4 פונקציות hash על מספרי הלקוחות:

לקוח/מוצר	1	2	3	4	f_1	f_2	f_3	f_4
1	*		*		1	4	7	5
2		*		*	2	1	6	7
3	*		*		3	6	3	3
4	*	*			4	7	4	2
5			*		5	3	2	1
6	*	*	*		6	5	1	4
7		*		*	7	2	5	6

א. חשבו את ה- Jaccard Measure בין זוגות המוצרים (1,2) ו- (2,3)

ב. חשבו את חתימות ה- Min hash לכל אחד מהמוצרים (לפי כל אחת מהפונקציות).

ג. חשבו את מידת הדמיון כפי שתתקבל מאלגוריתם Min hash בין כל זוג מוצרים.

ד. אם נבצע Locality Sensitive Hashing (LSH) עם שתי רצועות (f_1, f_2 ו- f_3, f_4), האם קיימים זוגות מוצרים שבוודאות יהיו מועמדים, כלומר שיש להם לפחות רצועה אחת זהה? אם כן, אילו?

ה. האם ייתכן מצב שבו מידת הדמיון שמתקבלת ב- Min hash תהיה קטנה יותר ממידת הדמיון בפועל? אם כן, תנו דוגמא. אם לא, יש להוכיח.

ו. האם ייתכן שזוג מוצרים לא יהיה מועמד ב- LSH אף על פי שיש לו מידת דמיון רבה יותר מאשר לזוג אחר שכן מועמד?

שאלה 4

יש לבצע clustering לקבוצת הנקודות הבאה לפי אלגוריתם k-means תוך שימוש במרחק אוקלידי (L_2) עבור $k = 3$. יש לפרט את שלבי ההרצה ואת ה- clusters המתקבלים.

הנקודות: $\{(2,20),(2,5),(8,4),(5,8),(7,5),(6,4),(1,2),(4,9)\}$.

יש להתחיל עם המרכזים ההתחלתיים הבאים $x_1 = (2,20); x_2 = (2,5); x_3 = (8,4)$.

שאלה 5

הטבלה הבאה מכילה נתונים לגבי קונפיגורציות של מחשבים בהם פגע ה- worm לאחרונה וכאלה שלא נפגעו:

מערכת הפעלה	דפדפן	תוכנת דואר	נפגע?
Linux	Firefox	Thunderbird	לא
Windows XP	Internet Explorer	Outlook	כן
Windows XP	Firefox	Outlook	כן
Windows XP	Internet Explorer	Thunderbird	לא
Windows 2000	Internet Explorer	Outlook	כן
Windows 2000	Firefox	Outlook	לא
Windows 2000	Internet Explorer	Thunderbird	כן

א. יש לבנות עץ החלטה ע"י שימוש באלגוריתם ID3 על מנת לקבוע האם מחשב בעל קונפיגורציה מסוימת יפגע מה-worm או לא.

ב. בהסתמך על עץ ההחלטה שבנית, יש לקבוע האם מחשב בעל מערכת הפעלה Linux שמריץ Firefox ו- Outlook ייפגע מה-worm. ומה לגבי מחשב בעל Windows XP שמריץ Firefox ו- Thunderbird?