ניסיון גישור

גיבוי פיזיולוגי

אמו סמי

כולם יודעים שסיבוכיות של מיון של $O(n \log n)$. זה נכון אל פרם. זה נכון
שכל אחד מאיתנו לא באמת מדקן את זה אל פרם.
הדברים שאותנו נעשה בקורס הזה.
יש תיאוריה של מה שאפשר לחשב וכמה זה עולה וזה נעשה
מהקשר של מודלים אבסטרקטיים. למשל במקרה של מיון-
סופרים רק את פעולות ההשוואה ובכל זאת מתייחסים
לזה שגם צריך למשל להביא דברים ממקום למקום.
ישנו כל מיני מודלים (טיורינג, Random Access Machine
ועוד) שמשקפים את המציאות בצורה כזו או אחרת.
ההבדקה היא שאותנו מעניין את המתשבים ואותנו יודעים
מה התכונות שלהם. אבל האם זה נכון? האם אותנו
מבינים ב.יסוד ויודעים לתאר מודל מדויק? למשל,
אותנו לא יודעים במדויק את הטופולוגיה של האינטרנט, למרות
שאותנו יודעים אותו...

מדע מתבסס על תצפיות. מייצרים מודל. עושים תחזית
בודקים את התחזית. זה יכול לעזור לשפר את המודל
או שאולי נגלה שמשהו לא נכון. זו השיטה המדעית.
הניסיונות יכולים או לאשש או לסתור את התיאוריה (כי
ההבדל בין תורה מדעית לדת). תיאוריה שתיאי לא בסתר
מוות היא לא תיאוריה טובה. ניסוי טוב הוא ניסוי שמראה
לנו אישור שאנו יודעים קודם. כמו בדוגמה ההתנגדה - טוב
טוב הוא סוס שנכשלנו בו. עוד אישור חשוב בניסוי הוא
שאפשר לחזור עליו ולשחזר את התוצאות.
את צורת המחשבה הזאת אותנו רוצים להפעיל על מדעי
המחשב.

מה המטרה שלנו? למה ניסויים לה טוב?

- להכיר קצת יותר טוב ובאופן פרקטי את העולם שלנו
- איך אנו משתמשים את התיאוריה
- תמיכה בנושא של פיתוח והגדרת תחומים חדשים

<u>דוגמה:</u> לוקאליות

כולנו יודעים שמה שנמצא במסים caching וכו' לה הליקאליות של תוכנות. שאלה מעניינת היא איך למדת באופן מדויק את הליקאליות הזאת. איזו תוכנית יותר ליקאלית מבין שתיים? התשובה היא לא טריוויאלית. כמה אנשים חשבו על הנושא הזה.

- הדרך הכי פופולרית היא stack distance.

מסתכלים על רצף הכתובות בזיכרון שהתוכנית ניגשת אליהם. כל פעם שניגשים לכתובת יש שני אפשרויות - או שאנו אותה קודם ואז היו רבר אופשרי במחסנית שלנו אף שצאת הפעם העומק שהה ונעבור אותה למעלה. או שזו הפעם הראשונה ואז פשוט נכניס את הכתובת למחסנית. תכנות עם הרבה ליקאליות מעיקר תיגש לכתובות בקרבה הקרוב של המחסנית.

אבל אפשר להסתכל על העומק המאוצע וכל שהעומק נמוך יותר הליקאליות גבוהה יותר. אם הגישה אקראית לתלוטין אבל העומק המאוצע יהיה כמעט גודל הזיכרון.

מעשית, ברוב התוכניות המאוצע נמוך יחסית

- אוטוקורלציה - נסתכל על הגישות לזיכרון כווקטור של מספרים. נציג אותו ב- a ונעשה קורלציה מנורמלת

הכתובות: $a(1), a(2), \ldots, a(n)$

המאוצע: $\bar{a} = \frac{1}{n} \sum a(i)$

את הקורלציה נעשה בין $a(i) - \bar{a}$ ל- $a(i+1) - \bar{a}$.

אם זה אקראי סביר שהקורלציה תהיה אפס

ואם יש לקאלות אז הקורבן יכול להיות גבוהה כי
אז שתי גישות קרובות נוטות להיות שתיהן מקום הממוצע
או שתיהן מתחת ואז המרחק הוא יותר (כי שני הגבולות
שני סימן) ול המתחברים רק מגבולים אותה.
בגישה איקראית השלישית נאבדו את החיבים.
• מספר הקומבינציות שמופיעות - למשל, תמונה בכל הצורות
הקוקבים אז השלישית רק הזאה ונסתכל כמה פעמים מופיעים
צירופים ראלה. כל שיש יותר הופעות הלקאלות גבוהה ביותר.

מה הקשר בין הגישות? אנתנו מקווים שאם גישה אחת
אומרת שיש לוקאליות, אז גם האחרות. אבל מעבר לזה
נדרשה לדעת דברים יותר מסובכים - מה הדיון של שיטה
מסוימת אומר על יעילות ה- cache? אם נעשה את
ההפיקה בדורה וסודות אולי נוכל לתכנן cache יותר
טוב וגם תשובה!

<u>בוחמה</u>: תקשורת עם דימיון לזמני
במשך הרבה שנים נותות של רשתות תקשורת (זומס,
buffering) התבסס שמתק שורת היו תהליך פיסוסון.
התקשורת עוברת בתחבולות ותשלוח היא איך הן מסות -
מה הרווחים בין התחבולות. ושם תקשורת בדרך מתתולה להיכנס
לדרור ושתנידות מגיעה ל-ווסט, אז בין התחבולות יש
רווחים גדולים ותשלוח היא מתי מסה התחבולה הבאה.
בתהליך פשאסון אנחנו אומרים שהתחבולות מסות בתפיגות --
אחידה בזמן. הבעיה היא שהלא מתשקרה במציאות.
בתחבולות שנות ה- סף אוספו רישומים מאוד מדויקים ומהר
מאוד הגיעו למסקנה שלה לא אחיד ואוקי זמן - יש הרבה
bursts ויש הרבה זמנים שאין בהם פעולות. יש
הבדלים שיקרים ברצוליות שנות ולא בסקאלה אופיינית
אחת. במציאות יש גבולים מה מאומר המקרב.

המדיפה מראות שיש דמיון עצום. אם מסתכלים על המדיפה מרצולוציה שונה היא נראה אותו דבר !! (פרקטל)

זה לא קורה בהתהליך פיטיטון. יש זה לא מיני השכות. קודם כל זה אומר שיש תלות ארוכת טווח. מה שקורה עכשיו יכול להשפיע על מה שקורה בעוד שניה אפילו אם על מה שקורה בעוד שעה ולכן יש השפעה גדולה על הגדלים של החולדים שאנחנו בוחרים, כי הדברים כבר לא מתמדעים בדרך אחידה, אלא על הזמן יש פלקטואציות שלא מתמדעות בשום סקלאה.

<u>דוגמה</u>: הקשר בין תיאוריה לניסיון - בעיית קיום שמות ועוד.

k-SAT - יש נוסחה בולינרית בדרך נורמלית עם n משתנים עם m חלקים ו- k משתנים בכל חלק. השאלה היא אם קיימת השמה מספקת - האם יש השמה למשתנים כך שיוצא true?

$$(x_1 \wedge x_3 \wedge x_5) \vee (x_5 \wedge x_4 \wedge \bar{x}_{12}) \vee \dots$$

אם $k=1$ זה מאוד פשוט. לוקחים את החלק הראשון והופכים אותו ל- true.

אם $k=2$ עדיין קלת תשובה גם אפשר למצוא השמה מספקת.

אם $k\geq 3$ הבעיה היא קיום-שמה. תיאורטית זה אומר שבגדול אין בריה אלא לבדוק את כל ההשמות.

<u>אמן</u>, רוב המקרים הם פחות קלים! בהינתן נוסחה אקראית אפשר להפוך שני כלי אבחנה:

- לנסות למצוא השמה מספקת: קבלה שרירותית של משתנים + back tracking
- לנסות לבדוק שאין השמה מספקת: שנמצאו סתירה.

מתברר שהסיכוי שהנוסחה טפיקה תלויה ב- $\frac{m}{n}$. ברור שאם יש מעט חלקים והרבה מאוד משתנים אז המשתנים מופיעים בהרבה מקומות ומשפיע על הרבה דברים וזה קשה ואילו אם כל משתנה מופיע בערך כמה מתוך מאה אז זה קל.

הניסויים מראים שיש סל מסוים בין ההתנהגות השונות. יתר מזה, כל שקרובים לסל הזה יותר קשה להחליט אם יש השמה מספקת או לא. סומר דריך יותר זמן רבי להגיע לתשובה. תרפים במעבר

הצעדי הראשוני שלה היה ניסיוני, אבל ברגע שעברו את זה התיאורטיקנים התחילו לחקור ומדשין יש תיאוריה והוכחות וניתוחים מנושאי הלה.

אבל גורם דומה לאיך הניסוי עבר לנו - אותנו יודעים מה דרוש להיות התוצאה אבל נשאל רבי להוכיח ולחקור אותה. המרוץ הוא שאותנו יודעים מה אותנו מחפשים.

ההבדל בין סיבות של מערכות הדשות לניסויים:

סביר שאותנו ניתקל ברעיונות שעל הניר נראים מעולה אבל במציאות לה לא עובד כל כך. ואז נוכל מעבר של עיבוד מימוש ובדיקה. דוגמה טובה לכך היא Planet Lab

ניתן שהמבנים פרוטוקול חדש ומגניב. אותנו רוצים להוכיח שעובד באינטרנט - סומר על רשת אמיתית ובגדולה שיש בה כל מיני הפרעות שאת קשורות אלינו ומשתמשים אמיתיים ותעבורה אמיתית ואת כל זה אותנו רוצים לעשות בלי להרוס את האינטרנט שקיים ועובד כיום. אבל Planet Lab זו מעין רשת וירטואלית שמופעלת על הרשת האמיתית. כרגע להיום יש להם כ-1000 מחשבים שמתנהגים כמו רשת אמיתית אלא שחלק מזה שהם משפיעים על התעבורה הם לא משפיעים על שאר הרשת. אבל אם המבנים שיפור ל-TCP אפשר לבדוק אבל שם, בתנאים הכי קרובים שאפשר לתנאי אמת

מה הפעולות שעושים בהקשר של מחקר ניסיוני?

- תכנון ומדידות
-
- חזרות על ניסויים

תלבטים ומדידות:

קודם כל צריך לאסוף מידע. כדי מבלי לא פשוט בהכרח. אחרי שיש מידע צריך להסתכל עליו ולהבין אם הוא שמיש, מייצג ואמין. יכול להיות שיש לנו מידע לא טוב- אולי כי הוא לא קרה בכלל ואולי כי הוא תלוי מאד מנקודת אמנו. לא בהכרח צריך להתעלם מדברים חריגים אבל אם אנחנו רוצים לבחון סיבות שאת- צריך לדעת לזהות את הדברים החריגים.

לסוף נמדוד את הדברים שנראה למדוד וכה שזה לא טריוויאלי בכלל. צריך להחליט מה למדוד ואיך ברגע לעשות זאת. בשביל זה צריך גם שיהיו לנו מטריקות למדידה. למשל, נתנו כמה קופונים למדידה של לוקאליות. יש דברים שבכלל לא ברור איך למדוד. למשל, אם יש לנו כמה מכונות וירטואליות על המחשב נראה שהן תהיינה כמה שיותר מופרדות אחת מהשניה. איך אפשר למדוד את ההפרדה? לא ברור...

בהנחה שאנחנו יודעים מה למדוד ומה המטריקה, נשאר לנו הענין הימראני- בדר"כ עדם המדידה תשנה לנו משהו. יש מקרים מודדים שבהם אפשר לקבל מידע כללי להשפיע על המערכת אבל זה ממש מיעוט. בדר"כ המדידה משפיעה על המערכת וגם צריך לוודא שאנחנו מודדים את כל מה שצריך כדי לקבל תמונה מלאה. חשוב מאוד גם לרשום מדויק את התנאים שבהם התבצעה המדידה.

אחרי שרבו עשינו את כל המדידות צריך אירשאן להליך את התוצאות בצורה ברורה. בדר"כ עושים את זה עם גרפים (בצורה בצורה גם עוסקים שני התרגולים הראשונים). להציג נתונים מורכבים ממש אומנות! לסיכום, נחזיק את הנתונים עם אחרים (כדי שנוכל לתקף עוד)

מה זה הפעולה?

- מודל שמסביר את התלפיות
- המודל מאפשר לעשות תכניות
- המודל יכול להיבדק בצורה ניסויית - לאשש או לסתור אותו

תכונות של ניסיונות:

נניח לרגע שהדבר הכי חשוב באחדות הוא שאפשר לחזור על ניסויים ולבדוק את התוצאות נכונות או לא. בדיוק לזה רמובן לא קורה כי אי אפשר לפרסם את אותה התוצאה פעמיים...

הדבר החשוב באפשרות לחזור על מדידות הוא שאנחנו מבדוק מהם התנאים המדויקים שבהם נעשה הניסוי. ובדרך שאנחנו מנסים לשחזר את הניסוי אנחנו מגלים מה חשוב ומה לא - האם גרסת ה- driver חשובה או לא? האם מרחק ההפעלה חשוב או לא? לה מקדם את ההבנה של מה שקורה במעבדה -

עוד דבר חשוב הוא שאפשר לחלק את העבודה בין קבוצות מחקר ובדרך לקדם את המחקר מהר יותר. למשל:

- TREC - חברות של חיפוש מידע אפי נושאים. קבוצות שונות עובדות על אוסף מסמכים נתון ואמורה למצוא את המסמכים שקשורים לנושאים מסויימים נתונים. רק כולם עובדים על אותו המידע ואת אפשר להשוות בין התוצאות.
- תחרות פי רבה אוטונומיים של DARPA - המטרה היתה לשפר בעיקר את הצד האלגוריתמי ולא בשביל התחרות וההשוואה חשוב שהתשתית תהיה זהה לכולם. את DARPA ניתן לזיהה כביכול מנועים רובוטים עתיים ואותם נתנו לקבוצות המחקר שעבדו על האלגוריתם.

שבוע הבא - מרפיקה סטטיסטית

תרגול ראשון נמצא באתר.

⑤ 24/3/09
אישור
נסיון יומי

את התרגילים יש להגיש עד יום שני בערב ובמקום שלישי מדברים עליהם בכיתה. לכן - אין דחיות! אם יש מלואים או משהו כזה אפשר לקבל פטור.

התרגיל הראשון - להבין שני data sets של רותם פס בתור בסיס ההוכחה. אם מצליחים את הנתונים בסקאלה לינארית בקושי אפשר לראות משהו אבל רצוי להבין בסקאלה לוגריתמית. ראוי לציין גם שרשמצליחים במלבן רצוי להגדיל את הגודל של הפונטים.

פתרון אחר הוא לעשות שני גרפים - אחד zoom in - להסתכל על הקטע הרלוונטי בהגדלה. לפתרון הזה יש חסרון אם יש סקאלה לינארית היא אוטומטית יותר להבנה. יתר מזה, בסקאלה לינארית ניתן לראות את הצורה האמיתית של הגל.

נקודה לחשיבה - האם רצוי לחבר את הנתונים בקו או לא? יש טיעונים לראן אבסון. הנטייה שלי דווקא היא כן לחבר בקו אבל מי שמעוניין יכול לחבר בקו מקווקו או תלוי יותר אז משהו - שלא יוצר הרושם שהקו אכן מייצג את הנתונים האמיתיים.

משהו לחשוב עליו - להוסיף grid אופקי או לא? רצוי להוסיף אם זה מעט עוזר להבין משהו. למשל במקרה שלנו כשהרותם פס מתיילם מעט רואים את זה יפה אם יש grid. בהתשב לא יש לתת את הדבר עדיף לדפיסות של ה- grid.

תשובה: בדיר ה-A היו הבלתי תלוי ובדיר ה-y הוא התלוי! קשה לתשב על דוגמה שבה רצוי שלא יהיה הפוך.

אם יש כמה גרפים של אותו grid רצוי מאוד ששניהם יהיו באותה סקאלה, אולי במקרי קצה אפשר אחרת... אולי אם יש יתרון מדידה שונות...

בתרגילים הבאים כמעט בהכלל יהיו מלוכלכים גרפים ודרור מלקח אמינו לעשות אותם טוב.

היום נדבר על הלבד נתונים - מבוא לסטטיסטיקה, מלון שירות טכני. בפרט נדגיש את הדרוך והחשיבות בכל להסתכל על נתונים.

יש איזה סטטיסטיקאי שרתה מאומר בשנות ה. טד וילסון נגד שטיפת המוח שעושים לנו כל זה שצריך לעשות חישובים תיאורטיים ולהסתכל על נתונים ולצייר גרפים זה לא מדויק. אז היו הביאו ארבעה data sets שמסתכלות ושונים לגמרי אי אפשר להבין טוב. אבל מתברר שיש להם כל אני מדדים סטטיסטיים (ממוצע, סטיית תקן, רגרסיה לינארית וכו') זהים. אז לראות הינו מלכים שהם דומים, אבל אם מסתכלים על הגרפים רואים שהם שונים לגמרי!!

המסקנה: צריך להסתכל על הנתונים!!! צריך קודם כל לראות מה יש לנתונים להגיד. יש בחור בשם Tukey שרתה על זה שצריך מסוכן.

אז כשיש לנו נתונים אנחנו לא יודעים מה הם אומרים אנחנו רוצים להסתכל עליהם לעשות exploratory data analysis וגם כשאנחנו מגיעים למסקנות נרצה להגיע אליהן לאחרים. בשני המקרים אנחנו מעוניינים בגרפים מסוגרי עיניים.

במבחן יש דגמומים לכל אני גרפים שונים ומשונים.

מה קורה אם לא מסתכלים על הנתונים נכון? במציאות חלק יש מן אמיל כאלה שאוטומט את הטיפול ומוציאים מהם גרפים של השיפה לבאר לבדקים. עשו כל מיני מבחנים והגיעו לכל שמסודר לפי סדר הטיסות עם התקלות שקרו עם פירוט לפי הדגם של הטיל והחלק

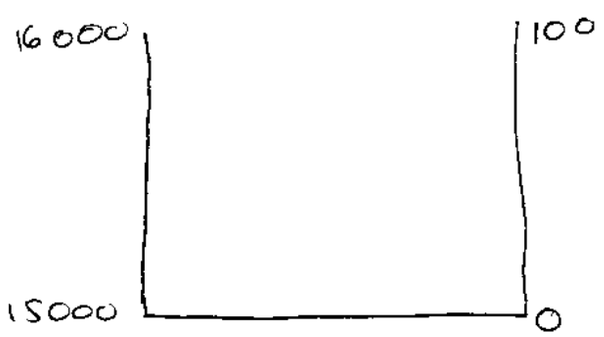
שבו קרתה התקלה. המטרה היא להבין אם תהיה תקלה בטיסה שמתבצעת בטמפרטורה מסוימת. בהנחה הכי שרוב האיכות בתרשים הוא לבן - לא אכפת לנו מה הטילים ומה הטיסה או איזה חלק היה תקול ובאיזה טול. כל מה שאכפת לנו ממנו זה מה היא הטמפרטורה ואם הייתה תקלה או לא..
בעזרת התוכנה, להסתכל והתלות התפוגגה (עם שבנה אנשי בוגר).

במצב - בראשון לכל אינני סומך אותם.

(7) 31/3/09
אישור
ניסיונות

הערות לתרגיל 2 – היה לנו אוסף שלם של data.
באופן מובהק חודש ספטמבר הוא outlier ולכן
הנתונים לו מעניינים.
- לא תמיד נכון שצמן התאמה וצמן ה wait איכותים
אור אותו הדבר
- כנראה עדיף להסתכל על השוואות bounded slowdown כי הם
לא שם דגש על עבודות קצרות.
- backfilling מומה לכולם וכן כמה עבודות נשארו בסוף
אבל זה לא מעניין להסתכל על זה.

- כשמדברים על אותו גרף נתונים בסקאלות שונות זה יכול
להיות מטעה מבחינת הה בבליים. למשל אם יש משהו כזה:



הבדל שם שני משבצות בדרך ימין הוא הרבה יותר גדול בהשוואה
מאשר הבדל של שתי משבצות בדרך שמאל.

- האם הבודקים תליים בסוגים? זה נראה מתבקש אבל
הנתונים מראים שלא. בעיקר יש בלגן ורעש ואין שום
דבר מובהק.
- לא לנסות התנהגות כללית מאוד שהמשה מחבר בין
הנקודות בגרפים.

היום אנחנו הולכים לדבר על מדידות. לתהליך המדידה יש כמה מאפיינים:

- הרצלוציה של מכשיר המדידה uncertainty
- הפיצור של המדידות - לא תמיד אם מודדים את אותו הדבר מקבלים את אותם תוצאות. precision
- דיוק המדידה - עד כמה התוצאה קרובה לערך האמיתי. אולי המדידה מושפעת מאיזשהם גורמים חיצוניים. accuracy

למשל, אם מודדים כמה זמן משהו במחשב אז המדידה עדיפה לוקחת זמן. אנחנו צריכים למדוא את הבעיות התשתיתיות האלה שהוציאו אותם החוצה מהמדידות.

כמובן, יכולים להיות גם תופעות רנדומליות שמשפיעות על המדידות בצורה רנדומית. על זה אנחנו נרצה להתגבר באמצעים סטטיסטיים. למשל, אם תוצרים של המדידה המון פעמים ושונים מאוצע זה יכול לעזור.

חוץ מזה יכולים להיות outliers שמלוכלכים לו מכיוון אופלו ואותם אנחנו נרצה לאתר ולהתעלם מהם.

מדידת זמן -
במחשב יש שלוש דרכים לספור זמן
- פסיקה של השעון
- time stamp counter
- שעת זמן נפסוטיקולי NTP כאשר השעה באמת יודע מה השעה האמיתית.

פסיקות של השעון - יודעות פסיקה במעבד וה- handler של השעון לי. מדיוק הפעלה יכולה להיות באיזו תדירות מגיעות הפסיקות והפסיקות עד זמן מגיעות מטיימר חיצוני.

המדידה סופרת כמה פסיקות היו. זה נותן רזולוציה גבוהה למדידה. למשל בלינוקס התדירות היא בערך 250Hz.

Time stamp counter - רגיסטר מיוחד שסופר את מספר המחזורים שעבר מאז שהמחשב פועל. אז זה סופר באופן מדויק כמה פקודות בוצעו אבל זה לא סופר באופן מדויק כמה זמן אמיתי עבר וזה משתי סיבות:

1) כשמחשב מתחמם הוא אוטו' מגר

2) במעבדים חדשים יש כל מיני אופטימיזציות שמורידות את מהירות העיבוד כשאין עומס כדי לחסוך בחשמל.

היתרון כאן היה שהפקודות נספרות במדויק והרזולוציה גבוהה.

- Network Time Protocol - לרוב המשתמש יש שרת NTP מיוחד שמקבל את הזמן מאיזשהו מקור סטנדרטי. כל הנodes ברשת מסתנכרנים עם השרת הזה. זה לא מושלם - למשל הזמן יכול לצוץ מאופן לא לינארי (בעל הסנכרון).

בלינוקס יש פקודה gettimeofday שהמשתמש שלה הוא במיקרו-שניות אבל לא כל המימושים נותנים את זה.

אנטומיה של מדידה

המסירות: אנחנו רוצים למדוד איזושהי פעולה של המחשב ב-system level, בלוקשות דברים יותר מדי מסובכים.

בדרך אפשר לעשות משהו כזה:

```
t1 = gettimeofday
do operation
t2 = gettimeofday
```

$t = t_2 - t_1$

הבעיה היא ש-gettimeofday לוקחת זמן !! וזו טעות סיסטמתית.

עוד בעיה שיכולה להיות היא שהפעולה לוקחת קצרה, שהשעון לא התקדם בזמן הזה, או שזה פשוט קטן בהיסט לפני המינימום רגזומית שיש לנו.

אז אפשר לתזרם לפי התהליך בלולאה, כאשר אנחנו משמרים את מספר ההפרעות המינימום.

אופציה אחרת היא לשים רק את הפעולה עצמה בלולאה ולמדוד בתוך אבל אז צריך לקחת בחשבון את ה- overhead של הלולאה.

הבעיה הגדולה היא שהקומפיילר עושה הרבה אופטימיזציות שפוגעות לשנות את האופן שבו התכנית רצה. אז צריך לעשות פה בקרה ולנסות כל שניתן שהקומפיילר לא יאפטמז דברים התולה.

עכשיו נניח שהכל יש לנו מדידות. כמובן, יכולות להיות חריגות, אבל אולי זה לא סתם משהו אלא אולי הן מייצגות איזו תופעה שיש לה הסבר (דוגמה לגבל מאבגר- תיבה של 100bytes צריך בים בקומפל)

התרגול הבא שלנו הוא למדוד את הזמן של trap אבל לא מדויק כמו שעשינו בתרגול ב- OS.

חוות סמק - איך מודדים את הטווח שבו אנחנו חושבים שהעכבר האמיתי נמצא?

אנחנו מניחים שהמדידות שלנו מגיעות מהתפלגות כלשהי. בגלל התפלגות נורמלית ואז אנחנו מחפשים איפה נמצאת המסה הקריטית של המדידות.

המאוכלוסיה שבה ההתפלגות בו ההצרכה שלנו לערך האמיתי (אמוסף הנכון המכבי). אבל אם נעשה שוב לעם סט שם מדידות זה יכול לתת אמוצע קצת שונה. אז אפשר להסתכל על ההתפלגות של הממוצעים שהם לה כמויות תהיה בערך פזמון אבל הרבה יותר צרה. אז הממוצעים אפולגים סביב הערך האמיתי. ואם אנחנו יודעים את ההתפלגות אנחנו יכולים להגדיר טווח שבו נמצא 95% מהמדידות שלנו. וזה ה- confidense interval

כל המאמר אופיינה במצבת.

(10) 21/4/09
אשור
ניסוניות

ברוכים הבאים.

במעבדה - דיון בתרגיל 3.

Microbenchmarks

לפעמים רוצים למדוד פרט ספציפי במערכת, למשל כמה זמן לוקח לעשות context switch, קריאה לפס וכו'. את המדידות האלה עושים עם microbenchmarks.

היום נדבר כמה עבודות בתחום הזה.

lmbench

המטרה היא להתמקד בדברים הבסיסיים, אבני הבנין של המערכת. זה מאפשר לנו להשוות בין מס' שונות ולאפיין דברים שמשתמשים באבני בנין אלה. מה שמשנה הוא שזה פורטבילי. אבל יש שם כל מיני מדידות שעל פני השטח נראות מוזרות אבל הן נעשו כך רק שסביר יהיה להניח שזה עבודה של מיליון מחקרות. הם מדדו כל מיני דברים - כמן להבינו מהסיבה, כמן לזכור תהליך, כמן של trap ועוד. - פירוט במעבדה

mhz

מטרתו לגלות את קצב המעבד - כמה cycles היו עושים בשניה. לראות זה לא בעיה - לוקחים קוד שיש בו כמה instructions, מתקעים ומקבלים. יש פה כמה בעיות: - רבוגבית

- בקבלה אחת ב-C זה כנראה יותר מפקודה אחת ב assembler
- מעבדים מודרניים יכולים לבצע יותר מ instruction אחד ב- cycle

אבל אפשר לראות באסמבלר, אבל אז זה לא פורטבילי..
אז היה להם רעיון מאוד מגניב - בהנחה שם פקודה ליקחת
מספר שלם של מחזורים אפשר למדוד כל מיני דברים ואז
לנסות למדוא איזשהו מכנה משותף. - כדה שנכנס מספר שלם
של פעמים בכל המדידות.

(11) 5/5/09
גישות
ניסיוניות

תרגיל 4 - נמדוד רק בחלק הניסויים. נא לבדוק לה עתיק!

דברים מעניינים:

- אפילו אם רוצים לשנות ביט אחד, עדיין המחשב עובד בבלוקים. לכן הגדלים הראשונים לוקחים את אותו הזמן
- כמעבדים מודרניים אפשר להגיד על קטעי זיכרון ספציפיים עם לגשות Cache ואז באמת התוכנה עושה מה שאומרים לה ומקבלים תגובות אמינות. במקרה כזה רואים בבירור את ההבדל שבגודל שלגדלים פחות ממלה ויותר ממלה.
- אם מקבלים מדידה שחושבים שיש בה outlier ~~[illegible]~~ צריך לתת עליה עוד כמה פעמים כדי לוודא שזה באמת outlier ושאין שם שום תופעה קונסיסטנטית.
- KB יכול לסמן גם 1000 וגם 1024.
  KiB תמיד מסמן 1024. תראי להשתמש בזה כי זה מונח מלבבל.
- יש מקום אולי גם לבחון את הקרה בגדלים שהם לא חזקות של 2.
- אם יש לך מיני דברים שאותנו יודעים - גודל cache, זיכרון וכו' נראה לסמן אותה בגרף.
- רשרושים חושפים מהסוג הזה חייבים להשתמש ב- double ולא ב- float. רק אם יש בעיה חמורה של זיכרון ואותנו בטוחים שהדיוק של float מספיק, אפשר להשתמש ב- float.

ניתוח מידע

אז נניח שיש לנו מלא מדידות. צריך לנתח אותן איכשהו. אפשר לתאר את הנתונים - ממוצע, תצפיות, סטיית תקן וכו' ואפשר גם למדל את הנתונים. למשל, אפשר להתאים להם התפלגות, למצוא מודל שמתאר את מה שקורה - איך נוצרים הנתונים האלה.

איך יודעים מאיזו התפלגות הגיעו הנתונים? יש איכשהו סט של התפלגויות שהמידע יכול היה להגיע מהן. אז אפשר לשערך את הפרמטר של b אחת מהתפלגויות עבור הנתונים ואז לבדוק כמה ההתאמה טובה.

בדר כמו כן יכול להיות שאוטומטית אבל הבעיה היא שגם טוב רק להתפלגויות האלה שמוגדרות מראש ואי אפשר למצוא שילוב של התפלגות.

מה אופן החלק המעניין באן הוא מציאת הפרמטרים.

יש שתי גישות.

- התאמת מומנטים

דוגמה: בהתפלגות γ יש שני פרמטרים - β ו α ומתקיים שהממוצע הוא $\beta\alpha$ והשונות הוא $\beta^2\alpha$. אז אם יש לנו נתונים אמיתיים x, אפשר לבדוק את הפרמטרים מהנתונים האמיתיים:

$$\alpha = \bar{x}^2 / var x$$

$$\beta = var(x) / \bar{x}$$

זה טוב כי אם יש מספיק תצפיות אפשר לחשב את הממוצע והשונות ולקבל קירוב טוב. $(M_k = \frac{1}{n}\sum x_i^k)$

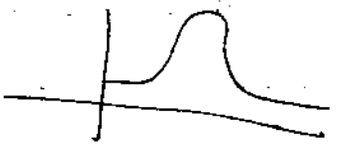
אז אם יש k פרמטרים צריך k מומנטים וכן בעיה בה מומנטים גבוהים יותר כישים יותר לרעש.

זה בעיקרי במיוחד להתפלגויות לא נגד כמה.

- Maximum Likelihood -

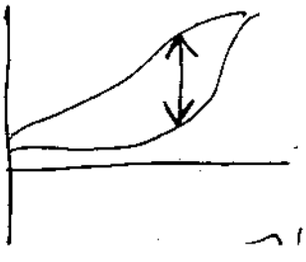
הרעיון הוא למצוא את הפרמטרים שהיו הכי סבירים שהובילו לתוצאות שנצפו. דומה מאד.

גם מקרה, כמו אחרי שמצאנו את הפרמטרים הכי טובים זה לא אומר שההתאמה הזו טובה. למשל, אם הנתונים הם



אז שום התאמה אקספוננציאלית לא תעבוד!

קולמוגורוב - סמירנוף - בודק כמה ההתאמה טובה. המטריקה היא המרחק האנכי המקסימלי בין ההתפלגויות



הם אומרים שכמה מדד לדימיון יש תיאורית שלמה של הנושא הזה. יש טבלאות שמתשבות בין מקל הדגימות לבין המרחק המקסימלי המותר. כלומר שיש יותר דגימות מצפים שהמרחק יהיה קטן יותר.

הבעיה הבאה לא כך מתאימה אם הצבות תשובים לנו והבעיה של Anderson-Darling יותר טובה למקרה הזה.

Q-Q Plots - גישה יותר גרפית בזיקה.

מודלים גנרטיביים -

אם יש לנו נתונים אפשר לנסות להתאים להם את כל ההתפלגויות שמצאנו במספר סטטיסטיקה, אבל רגעו לרגע את טווח התיפוש להתפלגות שאנחנו חושבים שיש הבדלים שיתארו את התהליך.

מיצוי מצונזר

בדיון שאותנו יכולים להיתקל בה היא שחלק מהנתונים "מצונזרים". למשל אם אותנו מודדים זמנים של מיני דבדים יכול להיות שמשהו התחול אבל עוד לא נגמר. אז יש לנו רק מידע חלקי - חסר תחתון על הזמן.

אבל אותנו מעל נאות כיכולים לבדוק ריכוב. אז אם ה 95% מהמידע חלוי בנראות אפשר פשוט להתעלם מהמצונזר, אבל אם זה חלק חלק אז אי אפשר להתעלם כמובן.

נסמן ב- x_i את הדגימות כולן.

d_i - מספר הדגימות האמיתיות (לא מצונזרות) עם הערך x_i

n_i - כל הדגימות (אמיתיות ומצונזרות) שהן לפחות x_i.

אז הסיכוי להישרד עד x_i ואז למות מיד הוא $hazard = \frac{d_i}{n_i}$

והסיכוי לשרוד עד אחרי x_i הוא $1-\frac{d_i}{n_i}$

מכאן אפשר לשחזר את פונקציית ההתפלגות האמפירית:

$$Pr(X \geq x) = \prod_{x_i < x}\left(1-\frac{d_i}{n_i}\right)$$ Kaplan-Meier Formula

ההסתברות לעבור עוד מספיק

הוא ההסתברות לעבור את כל אלה שלפנינו.

זה עובד בעיקרון מובן - נתונים בלי צורה. ההנחה רק היא שהאוכלוסיה הומוגנית (כולם מתנהגים באותו אופן) ושבכל שתיכנו יותר זמן יש סיכוי גבוה יותר להפסיק (זה נאיבי שכל יחידת זמן מגבילים מטבע שאומר אם להפסיק או לא ולכן ככל שעובר יותר זמן יש סיכוי גדול יותר שהמטבע שם יגיד לא להפסיק).

(13)

אך ארשין דיברנו על מדידת פריט אחד (למשל הגובה, מודל קבל וכו'). אבל לפעמים יש יותר ממדד אחד. למשל האם יש קשר בין גיבורון שתלוך דורך לבמן שמוא רל? אספר הגישות לקבל תלות באמל פלו...

במקרה כזה אפשר לעשות scatter plot ולשאלה היא האם מראים איזשהו pattern ואם אתנו יכולים למדל / להבין את זה.

הדרך לעשות מודלים כאלה היא ע"י רגרסיה.
המודל הכי פשוט הוא מודל לינארי $y = ax + b$. כשש
a ו- b פרמטרים שצריך למצוא.
במקרה כתלה אין מודאים אותם.
ואז שם צריך לבחור את איכות התוצאה.
ובכל מקרה, צריך להסתכל על הנתונים ולראות אם אין שם איזה outliers שגרמו לנו את הזינים.

(14) 12/05/09
אישור
ניסויים

הערות לתרגול היום -

- אם יש נתונים שאותנו לא יודעים מה המשמעות שהם (phantom) עדיף להתעלם מהם.
- לא להתעלם מערכים גבוהים - יש מצב ראלה אבל זה לגיטימי... דרך סיבה טובה בשביל לבדוק נתונים!!
- יש הרבה דרכים של אפס ואותם דרך להדריך.
- אם מסתכלים על הגרפים של hang! - agent רואים שהם חוזים את זה ועד מראה לנו שהם פעם באוו מאוחרת ההתפלגות אלה שנמדדו היו יותר אחרים מושש. מידע שהשג לא מלמד דרך להיות יותר שמשלה ולמשלה מהמידע המלמד.
- הרבה הנסרע של הנתונים הוא מלמד וזה יכול בעיה לקירוב ואותנו הדברים שגורם לאתתק בין השירוק למדיאות, בעיקר בכלל.

אנחנו מדברים על הנמן של מדעיות של מערכות. דיברנו על הטכניקה של המדידה. אבל יש גם שאלה ברמה יותר ברמת גבוהה יותר: אם יש מערכת עם רמה קונפיגורציות, אז מה למדוד? ואיך ללמוד משהו על הה שוואה?
למערכת יש גורמים (מעבד, זיכרון וכו') ולגורמים יש רמות (סוג המעבד, כמה זיכרון וכו'). התכנית היא איזה דירוג של גורמים ורמות שונים למדוד בכל ניסוי.

Simple factorial design - מתבוננים מקונפיגורציה בסיס ואז לוקחים את כל הגורמים בנפרד ומשנים אותם. אם יש k גורמים לכל גורם n_i רמות אז מספר הניסויים הוא $1+\sum_{i=1}^{k}(n_i-1)$.

כדי מראה יותר ההשפעה של כל אחד מהגורמים אולם גם מפסיד יותר האינטראקציה בין הגורמים.

- full factorial design הם יכולים ההפורה היא לבדוק את כל הצירופים האפשריים של כל הגורמים ואז צריך $n\prod_{i=1}^{k} n_i$ ניסויים וזה יתכן להיות הרבה מאוד עבודה.

fractional factorial design - דרך הנהה - מודדים קבוצה חלקית של הצירופים. כמובן התלך הקשה הוא לבחור איזה צירופים למדוד.

- design - 2^k - מודדים את כל הצירופים של הגורמים בשימוש בדרגות הקיצוניות שלהם.
- design $2^k r$ - 2^k רשחוזרים שלם מפורה r פעמים.
- design - 2^{k-p} - בוחרים רק חלק מהצירופים.

בדוגמה הבאה פשוט הוא 2^2 - יש שני גורמים שלכל אחד שתי רמות. הרמות בדרכות לרוחב כלומר וסמירות. זה כמובן רחוק להשתנות עם הזמן / טכנולוגיה.
מכל אמצע, דרך למדוד ארבעה צירופים ואז דרך לנתח מה המשמעות של זה.

נגדיר מקדמים מוגדרים עם ארבעה כמטרים q_0, q_A, q_B, q_{AB}
$y = q_0 + q_A X_A + q_B X_B + q_{AB} X_A X_B$ - התוצאה
X - הרמות של הגורמים ; q - כרמות שרוצים לחלוק.
נסמנים שרירותית איזה הרמות של X ב- ± 1
ואז מהסתכלות בתוצאות הניסוי שלנו מתקבלות ארבע משוואות בארבעה נעלמים.

$$y_1 = q_0 - q_A - q_B + q_{AB}$$
$$y_2 = q_0 + q_A - q_B - q_{AB}$$
$$y_3 = q_0 - q_A + q_B - q_{AB}$$
$$y_4 = q_0 + q_A + q_B + q_{AB}$$

למשל אם נתבונן באמצע נקבל $g_0 = \frac{1}{4}(y_1+y_2+y_3+y_4)$
ובאופן דומה אפשר לשחק איתם ולפתור.
ואז מתקבל מודל כלשהו. נשים לב שאותנו לא מתעניינים
לעשות רגרסיה אינטרפולציה או משהו. הגורמים רוצים להיות
עם אפס וסיגמא - אותנו לא משתמשים ברמות בזמן.
סתם נתנו לכם סימונים - ± 1.
אז מה המשמעות של הדבר הזה?
g_0 - הממוצע של המדידות
g_A - ההשפעה של גורם A
g_B - ההשפעה של גורם B
g_{AB} - ההשפעה של האינטראקציה של A ו-B
תיאור גרפי במציאות.
דוגמה במציאות.

אם ההשפעה שהם נמוכה סביר שהם יחסית בלתי תלויים.

אבל למה דווקא מודל תיבורי אומר כלום?
יש מודל אלטרנטיבי כפלי שבו

$$\ln y_i = g_0 + g_A X_A + g_B X_B + g_{AB} X_{AB}$$

איך בוחרים באיזה מודל להשתמש? הדרך טובה להבין את המערכת
להתשובה אם המודל דרוש להיות כפלי או חיבורי. אם אין לנו
אינטואיציה אז לפעמים אם במודל החיבורי יוצאת השפעה
גדולה ל-g_{AB} אולי כדאי להתחיל את המודל הכפלי.

אז ערשיו בואו נראה דוגמה שיש שני גורמים עם שתי רמות
ועשינו חבילת אלטאה. אבל אם יש הרבה גורמים נגלה
לעשות אבניות תלויות ואז דרוש אפשרות לבחור את הדרגה.
מה גדול ההנחה היא שכמה שאינטראקציה מסדר יותר גבוה
ההשפעה שלה יותר נמוכה.
הסבר ודוגמה במציאות.

(16) 19/3/09
אושות
נסיון יב

הערות לשעורי בית - תרגול 6

- אם מתחברים קוד מתרגלים קודמים זה בסדר אבל צריך לעשות הכל בראש. הרעיון הוא לישם טורמות שלומדים מהתרגול ולא להתעלם מזה לגמרי.
- בגדול היה מאוד רחב של תוצאות ולא ברור מה הסיבה - אולי באלגו סביבות זהובה שונות ואולי כי במטלות הבדרים מאוד מסובכים וירגל להיות הבדל בין קומפיילרים שונים.
- האינטואיציה אומרת שמודל חצלי הוא טוב יותר כי המחשבה איך שמו קובעת את גודל הנתונים שצריך לטפל בהם. צריך להסתכל על המערכת ולהבין אותה ואז לבחור מודל.
- רשמתקבלים הפרשים קטנים בין המקדמים צריך לחשוב אם זה באמת בגלל משמעות וירגל להיות שהם היה שינוי קטן בזרים של הפרמטרים, הבחסים היו משמעותיים.
- צריך להסתכל על התוצאות ולנסות שמן הגיוני. יכול להיות שקבה משמו מהלכים המדידה והיא השתבשה. אם התוצאות מוזרות - צריך לחבור על המדידות!! אז אולי חלפין אנחנו בלב לא מבינים את המערכת ואנחנו מודדים את הפקטורים הלא נכונים.
- המסקנה הבאלה היתה שאין מסקנה - זו תמיד אחד המורכמים חשוב יותר.

Workloads

בארגו לומדים שלמן הרילה חלוי מאוד הקלט מדב"כ ומתבים על המקרה הגרוע. רשאנתחים מן הקלט הוא הגמס-

הפורים שנותרים למד לעשות ומעניינת אותנו הבידוקים בממוצע.
אבל הינו חבים לאמץ את הנומסים ומשמעות הערכת בידוקים סטטיסטיות.

בקורס הזה אנחנו נסתכל על נומסים אמיתיים וננסה לאמץ אותם.
מבוא המהרים המד' שאנחנו מסתכלים עליהן הן הרחבה של מד' ביטוחות ומכן אפשר לקחת נומסים של המד' ויקיאות ולבדוק אותם
את המד' החדשות ואז אפשר לאסוף את המידע אקטיבית (לא של לשים sniffers לרשת) אלא להשתמש במידע שקיים כבר.
למשל, הרבה מד' אוספות מידע בשביל ניהול, רשתות web יש לוגים וכו'.

אז עכשיו יש לנו נתונים. מה עושים איתם? בדרך רוצים לייצר מודל סטטיסטי כלשהו - התפלגות של כל מיני מאפיינים והקורלציה ביניהם (האם תהלכים ארוכים משמעשים ביחד ביכרון?)

שתי תופעות נפוצות בהתפלגויות במתשבים
- התפלגויות נוטות להיות מוגלות - יש ערכים דיסקרטיים מאוד שכיחים וזה שביניהם נדיר - למשל יש איזו נטייה לעבוד בחבקות של 2 רשמרים עמודות מקבוצות. אותה תופעה קיימת בגדלים של קבצים ומאגרי תנועות ברשת.
- התפלגויות רוצות להיות עם זנב כבד - יש mass-count disparity - רוב המסה נמצא בזנב של ההתפלגות. למשל רוב התהליכים קצרים אבל את רוב הזמן של ה- cpu עסוקים תהליכים ארוכים. אם משרת web - רוב הדפים לא פופולריים אבל רוב הבקשות הן לדפים פופולריים.

תופעות של נומסים
- לוקליות. למשל אם עכשיו בא תהליך קצר אז סביר שאחרי יבוא עוד תהליך קצר. אז דאגה אקראית מתעם ההתפלגות

יכול מאוד בדרך אפיקת את הנתונים.

הדבר הראשון שדבר ראשון ושמתבלים נתונים הוא לנקות אותם - מבלבל ומנתונים שלא מעניינים אותנו. לצרוך נתונים שהם בכלל בה בסיטי כך בה שלוי לשנות אות והתפלגות. אבל דרך לחשוב על דברים ראויה טוב טוב שהבין מה המשמעות של בריקת נתונים.

במדעית - כל מיני דיאגרמות לנתונים שברור שהם לא מעניינים. אבל כדאי להתעלם מהם.

יוד יל בנה כבד -

כמו שאמרנו - יש הרבה דברים קטנים. אבל אולה שגדולים הם **מאמש גדולים**. וזה קורה בהרבה מדידות - אורכי תהלירים, גבולי קבצים, פעילויות של דפים. יי mass-count disparity - גול, להיות מתואר ע"י יי joint ratio, כלל שה- ratio גדול יותר זה יכון יותר.

הגדרה הפורמלית שב בנה כבד היא שהבנה בוצר פולינומיאלית - יש α כך ש-

$$\bar{F}(x) = Pr(X > x) = x^{-\alpha}$$

$$\Rightarrow \ln(\bar{F}(x)) = -\alpha \ln x$$

זה משמש גם למבחנים בנה כבד ויש לאפיון האינדקס α שלו.

מומר אם מבינים את הגרף לוגריתמי בשני הצירים אז יש יתם לינארי בבנה.

(18)
26/05/09
גישות
ניסים יניב

הערות לתרגיל 7 -

המטרה הייתה לתעד אותנו להסתכל בסקלטון של המידע שמתקבלים. מיעוט הנתונים היו בסדר, אבל היו רמה גבוהה משמעותיים שצריך היה לתת שליטה את הדעת

- היו קבצים בגודל $2^{32}-1$. מתוכם הרוב עם קוד 304 ו-6 עם קוד 200 לדעתי ה- 304 זה כנראה לא משנה כי לא ברור מה הדפדפן מסתכל שם בכל. אבל בדור שלדעתי עדיף עם קוד 200 כי ממש בעיה אם סמיר שבר נכון.
- היו שני משתמשים עם הרבה מאוד בקשות - הרבה יותר מכל האחרים. לפי הקבלה של הבקשות נראה ברור שזה לא יכול להיות בן-אדם חוץ מזה מסתכלים קצת יותר לעומק אבל נראה שהמשתמשים האלה בקשו את כל הדפים פעם אחת - נניח משהו שיש התנאים שמדובר ב-crawler או משהו כזה.
- יש דפים עם גדלים שונים במקרים שונים. יכול להיות שמדובר בדפים דינאמיים או אולי זה רק מידע חלקי (קוד 206).
- היו כ- 700 שורות בקובץ שתבו על זמן באופן מדויק. זה אמנם נכון מכל הכמות האדירה של המידע שיש לנו, אבל זה משפר שרדאי לשים לב אליו. יכול להיות שמדובר בבעיה ב- logging.
- 25% מהבקשות שהובילו לקוד 404 הגיעו מאותו האובייקט. זה מעלה את ההשערה שהיה לינק שגוי באתר. זה נותן רעיון לבדיקה אוטומטית של שרת web.
- כל הבקשות שקיבלו קוד 500 הגיעו מאותו משתמש. שאלה מעניינת היא מה הוא עשה שגרם ל- server error...

Workloads - המשך

השבוע נדבר על פרקטיקות ודימיון עצמי ועל הרחבה של מושג הלוקאלי.

בדרך אינטרוספקציה שעבודות מגיעות אל השפה בתהליך פואסון:

- העבודות מגיעות בפואסון בלתי תלוי אחת מהשניה,
- העבודות יכולות להגיע בכל רגע נתון בהתפלגות אחידה
- מודדים את הזמן בגרנולריות מספיק עדינה כך שבכל רגע יש לכל היותר בקשה אחת שמגיעה.

כה המודל הכי אהוב והכי חביב להגעה בקשות כי יש לו כל מיני תכונות נחמדות

- עבודות מגיעות מפואסון אחיד לאורך הזמן
  - אין פרצים של פעילות
  - אין מחסורים של פעילות
- מרווחי הזמן בין הגעת הבקשות מגיעים מהתפלגות אקספוננציאלית. (וזה מאפשר לסמלץ הגעות בקלות)
- איחוד של תהליכי פואסון יצא תהליך פואסון.

נאמר אם יש כמה אקורדים שונים שמהם מגיעות בקשות מתפלגים פואסון אז הכל ביחד זה תהליך פואסון.

- אם מסתכלים על תלון זמן יותר גדול הנשמרת יכולת - כמו שמסתכלים יותר זרמים ביחד, המלכות יותר גדולה וכו'.

לסיכום הדרך זה לא תמיד משקף את מה שקורה במערכות אמיתיות. אם יש לנו מידע אמיתי אפשר לבדוק אם הוא בא מתהליך פואסון אם מרווחי הזמן מגיעים מהתפלגות אקספוננציאלית. חוץ מזה צריך לבדוק שמרווחים עוקבים הם בלתי תלויים. אז זה אפשר לבדוק באמצעות קורלציה של המרווחים.

$X_1 \quad X_2 \quad X_3 \quad \dots$ הצאנים

$D_1 \quad D_2 \quad \dots \Rightarrow \bar{D}$ המרווחים

הקורלציה שמעניינת אותנו היא $\sum_i (D_i-\bar{D})(D_{i+1}-\bar{D})$
אם הם מלאי תלויים אז הממוצע יהיה קרוב לאפס.
לפעולת בדיקה לוודא שאריבה מוריד את השונות
בהיסטוגרמה.

מנתונים אמיתיים הרבה פעמים יש תופעה של דימיון עצמי.
האינטרנט לא מתלבט שום דבר ותעבורה נראית דומה
מאוד בכל סולמות שונות.

ולזה יש השלכות על תכנון המערכות. הפרצות (burstiness)
מאוד מקשות על הקצאת משאבים

אם מודדים דימיון עצמי המהות היא שיש קורלציה גבוהה בין
מדידות לקבוצות ואם מסכמים אז הסטיות יכולות להיות מאוד גבוהות.
אנשים היתה הרגשה מתמטיקה שמפורסמת במלחמה.

סוף לוקלי ות

מודל מקומי לייצור עומסים הוא בשימה מתוך התפלגות. אבל
יכול להיות שגם בהתפלגות מיוחדות אבל מסתפה עניין
הקורלציה, כי מדבר בשימה מהתפלגות מנתים אי-תלויים
ולכן לא מתבטא מאיפיין את המציאות. זה נעשה מבמה דברים:
- יש מתנודים יומיים ושבועיים וכו'
- ההתפלגות בפשוט משתנה עם הזמן
- בני אדם לא מתנהגים כמו מודל מתמטי - הם
עושים את אותו הדבר במשך זמן מסוים לאור הפעולה
משמעה.

בלוקלי מוריד את האינטראקציות ...

(20) 9/6/09
גישור
ניסיון

הערות לתרגיל ז -
הקבצים שהיינו צריכים להשתמש בהם הם בינריים ושם תוכנה שיודעת להמיר אותם למשהו קריא. אבל לאותשם התוכנה הזאת סבו מוגבלת לנו את כל הנתונים אלא רק ברזולוציה של 3 נקודות (מעוגל שבמידע המקורי יש רזולוציה של שניה).
אבל בגדול תמיד רצוי לבסתכל על המידע המקורי.

הערות על תרגיל ט -
- באמת הכיוונים חד-כיווניים או דו-כיווניים? גלוי אה אנחנו רוצים לה סיק אבל מבחינה פיסית בנראה שאם יש כבל בין A ל-B אז סביר שזה דו-כיווני ויש גם חשמל קישור בין B ל-A. מאידך, אם מישהו מ-A שלח לאתר ב-B זה לא אומר שגם הוא שקר מתרחש...
- מה אינטרנטיות:
  - לאינטראינה הייתה עליה מאוד גדולה בקישוריות ב-2009. אולי זה קשור לקבוצה מאוד פעילה.
  - לרוסיה היה איזה פיק בלינקים חיצוניים ב 2008.

ניסיון לזהות בני אדם

מעבר מחשבה מפוחדות ע"י בני אדם מתוך לקחת ע"י בני אדם אחרים מי שמשתמש בהן הם בני אדם. ושוטשת הקבוצות הגדולה לא בהכרח נתחבות. ייתכן דרוך הן לא בהכרח מבינות אתת את השניה... רק דרך לזהות ניסיון.
- הנחת תוכנה אמפירית - איך בני אדם מפתחים תוכנה ומהן שיטות עבודה יותר יעילות?
- מאשקי אדם מבחינה מבנו מאשקי עבודה, כמה תוג משמש?

שיטות וניסויים

- תצפיות וניתוח מדע - להסתכל על המשתמשים לראות איך הם מתנהגים (ללא התערבות)
- ניסויים מבוקרים
  - איך משתמשים מתמודדים עם סדרת פעולות מורכבות היטב
  - לראות איך ההתנהגות משתנה כאשר נעשה שינוי פעולפי במערכת
- רישומים וסקרים - לנסות להבין למה משתמשים מתנהגים כמו שהם מתנהגים. זה כמובן לא יוצא מהכלל מדויק אבל אם עושים מספיק סטטיסטיקה אפשר להבין לאיזה כיוון נושבת הרוח.

תצפיות - יש רישומים של פעולות משתמשים במערכת מתשבי מל. אז מסתכלים על פעולות של כל משתמש בפני עצמו. מגיעים שמשתמש עובד בצורה סדרתית - מגיש עבודה, גיבו מסיימת, הוא עושה הפסקה, מגיש עבודה וכו'. מתפשים קורלציה בין אורך העבודה לאורך ההפסקה

המסקנה: זמן ההפסקה גדל מונוטונית עם זמן התגובה (הזמן מתכנת שהעבודה הוגשה עד הרגע שהעבודה סיימה). ואולם בין ה slowdown להפסקה אין כל כך קשר (ה. slowdown זה היחס בין זמן הבילה ביחס של העבודה לזמן התגובה).

איך עושים ניסויים בני אדם?

- לא לבוא בגישה של ארכיון או מבקר אלא בגישה שהמשתמש הוא מנמה ואתנו באים ללמוד ממנו.

שבים בניסוי מבוקר -

- להגדיר את מטרות המערכת - מה הפעולות שהיא מאפשרת?
- ליצור אוסף מטלות שיש לבצע כדי להשיג מטרות אלה
  ההגדרה צריכה להיות מאוד מדויקת אבל לא להסגיר את
  דרך הפעולה.
- מדיאור קהל - למצוא קבוצה של משתמשים פוטנציאליים
  שמתאימה לקהל היעד ואז לסמן אותה ולהשאיר רק את
  אלה ששימושם.
- בפגישה זו המשתמש צריך להבהיר שהמץ לא מתנת
  אמנו אלא הוא עוזר לבתוך את המערכת. משהו שיציד
  מקל רמ מה הם עושים ומאה. כדאי אפילו להקליט את
  הפגישה. אסור בשום פנים ואופן לעצור למשתמש במשימה.
- מתן סיכום אפשר גם לעשות כיאין

(22) 16/6/09
אושרי
ניסיון

הדרכה לתרגיל -

- ירגל להיות שלא רוצים אנשים שבטל לא מבינים כלום - הרי אנחנו לא בודקים תוכנה לילדים בגן. אנחנו לא רוצים להתחיל אלה סביר מה זה זכר.
- צריך לבחור תוכנה שלא כולם מכירים אותה (למשל יש הרבה אנשים שלא מכירים Office 2007 או open office)
- אי אפשר לעשות סטטיסטיקה על מפות מאוד קטן אבל כן אפשר לקבל מצב מושג על מה שקורה. תוך מצב אפשר לראות כמה המשתמשים שונים זה מזה.
- אפשר לקחת מישהו כן מנוסה ולהשתמש בו כתנור קבוצת בקרה.
- רעיון נחמד זה להקליט את המסך והתנועות בזכר - זה יותר קל מאשר להקליט במצלמה ויד אז. זה גם אומר שרואים את המסך בצורה ברורה יותר טובה.
- תזכורת
  - כמעט אף אחד לא משתמש ב- help ומי שמשתמש לרוב לא מוצא את מה שצריך.
  - משתמשים מסוגים נעולים לאופלקציות של MS Office
  - ⇐ כדאי לעבוד את במסגרת דומה ל- MS Office
  - הרבה אנשים הופתעו מכל מיני התנהגויות שונות ומשונות של משתמשים
  - אנשים לא רואים את מה שיש להם מול העיניים
  - אנשים כל הזמן מנסים דברים מחדש אפילו שזה לא עובד
  - לפעמים אנשים עושים תפוש סיסטמתי של כל התפריטים
  - רוב האנשים חסרי סבלנות בדג'

- קשה לתלק משימות לקבוצות שונות של תפריטים
- יש נטיה לאנשים לדבוק על התפריטים שלמדה אבל מתמודדים מעולה שלמה.
- דברים שאנחנו חושבים לטריוויאלים לא בהכרח רואה למשתמש הלא מנוסה.
- אפשר לארגן דברים לפי הפעולה ואפשר לארגן לפי הדרך שלהן פועלים. ב-office 2007 דברו מסיוים לפי פעולה לסיוים לפי אובייקט ויש משתמשים מהתבלבלים שלנו שאנשים נוטים יותר לסווג לפי אובייקט.

דוגמא סיוים שאפשר לעשות הוא לפי פעולה לוגלית לדוגמת פעולה אלגוריתמית.

גישה לסיווגים לאלגוריתמים

הגישה הנשוטה בקורסים שעשינו היא מאוד מתמטית. מונים מודל, שמתרים בעיה, והם מתמטי יפה.
למשל, למדנו ש- Quicksort עובד בזמן $O(n^2)$ כאשר n גדל הכלל. נניח אם יש n אפשרים ליין האלגוריתם לעולם לא יפעל ביותר מ- cn^2 פעולות! זה טוב הסנוור הנתרכלות שאנחנו לומדים בקורסים באלגוריתמים למעשה.
בגישה זו כמה בעיות
- בדרך לא מדברים על הקבוע c. זה אומנם לא משנה כאשר $n \to \infty$ אבל בגדלים פרקטיים זה יכול להיות מאוד משמעותי
- בורות הנתונים הנראה מאוד קונסרבטיבית ואומרת לנו מה המקרה הגרוע. אי אפשר לעמוד מלה דבר על

המהדר האוניברסלי ומאז אפשר לכפות במעבר אמיתי

דוגמאות

- יש גרף ומתפשים על פורש מינימלי (MST) באלגוריתם של חוזק פותר את הבעיה ב $O(E \log V)$ מהדר שיש אלגוריתם הרבה יותר מסובך ומתוחכם שפותר את הבעיה בזמן $O(E \log^* V)$.
נשמע מטורף!!! אלא שמגאות הדי שבכל אפשר יהיה לראות שיפור דיבוג לחבוד על גרפים לפופים עם יותר מ- 10^6 קודקודים... אבל גם הקטנים מגאותיים ממש אין טעם להשתמש בהם.

- בדיקה אם גרף הוא minor של גרף אחר.
נאמר הוא minor של B אם אפשר לשנות את A ב- B אחרי שאחדו ב- B כמה מיני קודקודים).
אבל מתברר שיש לבעיה פתרון פולינומיאלי. מלבד שהקבוע הוא 10^{150}...

- רשת מיון היא מין ארכיטקטורה יעודית למיון.
שאלה: האם אפשר לעשות רשת מיון שבה (n פעולות שבמקביל) זו הייתה בעיה פתוחה הרבה זמן עד שמשני כ- שב שנה הראו שלא נתן. אבל גם ראו הקבוע הוא פשוט עונק.

- כשעושים אנליזה למקרה הגרוע בדר"כ זה מקרה נדיר ולכן האנליזה לא משקפת את המציאות.
הדוגמה הכי טובה לכך הוא אלגוריתם הסימפלקס שפותר את בעיית התכנון הלינארי $\max_{s.t} \langle c, x \rangle$
$Ax \le b$
$x \ge 0$

תיאורטית האלגוריתם הוא אלגוריתם שבמציאות הוא מתכנס מאוד מהר... יותר מזה, מתמטיקאים הוכיחו לעבור מאוד בשלב רפי למלא דוגמה לקלט שמצריך זמן הרבה שהאלגוריתם הוא אלגוריתם.

- יש מקרים שבהם אנחנו בכל זאת יודעים לעשות אנליזה. למשל בעיית מינימום צביעה של גיתוכי הקשתות כאשר מצבירים גרף דו צדדי. הבעיה היא קשה - שלמה. יש אלגוריתמי 3-קירוב עם נמוכים, עם הוכחות והכל. ויש אלגוריתם קירוב שיחס הקירוב בו לא קבוע ופם לא מעט יודעים להוכיח אותו שבמציאות נותן תוצאות יותר טובות מהאלגוריתמים המוכחים.

- האנליזה התיאורטית לא מתחשבת בהשפעות סביבתיות כמו כיבוי מטמון ואפקטים של עוקלות. אם מאוד בשלב לנתח דברים ראוי בצורה אנליטית מדויקת אבל יש להם השפעה עצומה על הביצועים.

מה באמת רוצים?

- להבנים תוצאות אמיתיות על קלטים מציאותיים ולא לשטות להציג את הקבוץ...
- להבנים מקרים נפוצים ולא מקרים נדירים מאוד שבגלל להראות התנהגות אמא נקודת מבדדה בוגדת

מתוגדולים יה

- לא לתקוף מימוש טוב ויעיל של האלגוריתם. זה לא מאמץ טוב.
- אם משווים שני אלגוריתמים שניהם צריכים להיות מומשים היטב וההשקעה לריכה להיות דומה - שתקיה האינות בסטטיסטיקה...
- הקלטים צריכים להיות יצוגיים למציאות קלטים אמיתיים.

- לבדוק את המאפיינים היומי - מה הם שציפרנו
לגיו בתחילת הקורס
- לדווח בצורה שלאה את התוצאות כדי שאחרים
יוכלו לשחזר אותה שנשיגו ולקבל תוצאות דומות -
זה מאשש את התוצאות.

(25) 23/6/09
שישי
ניסיוניות

הערות לתרגיל 11 - זמן ריצה של מיון

- המקרה הגרוע ~~שלו~~ של quick-sort הוא $O(n^2)$.
אבל אם יש לנו סתאום מדידה גבוהה איך נדע אם
הנתונים גרועים או שהייתה הפרעה? תוכנים של
המדידה עם אותן הקלט...
- באלגוריתם בקשה להחליט אם ההתנגשות היא לינארית
או nlogn. אפשר לנצל את התוצאות בשביל לראות מה עדיף
- ל- tsvipk יש השפעה על המיון...
- מה אומנו רוצים לעשות עם הה cache במדידות?
השאלה היא מה המטרה. כדי לתת גיבוי לתיאוריה אם
צריך איך שהוא לנטרל את כל עניין הירכיות הזיכרון.
אם כולם מבדוק מימוש אמיתי של מערכת מסויימת
אבל לא בהתנאים הסביבתיים יש משמעות גדולה!

Case Studies

איפה הגישה הניסיונית יכולה לתרום?

- online scheduling -

נתונה m מכונות ו- n עבודות והמטרה היא
להחליט איך עבודות להכניס על איזו מכונה כדי שהזמן
מתחלת העבודה הראשונה לבין סוף האחרונה יהיה מינימלי.
האלגוריתם הדמדן מאוד אינטואיטיבי - לכל עבודה שמגיעה
נשים אותה בתור למכונה שלפניה לוקחים הכי מוקדם.
אפשר להראות שהאלגוריתם הזה הוא $\frac{1}{m}$-תחרותי:
נסמן ב- c^* את הזמן האופטימלי שאפשר להשיג. נסמן ב- p_i
את זמן ריצת העבודה. ברור ש- $c^* \geq p_{max}$; $c^* \geq \frac{1}{m}\sum p_i$

נניח ש- k הוא העבודה האחרונה שהוקצתה לסיים. אבל בהכרח העבודה הזו התחילה לא המאוחר מ- $\frac{1}{m}\sum_{j\neq k} p_j$ כי אין מכונה שלא עובדת (העבודות מגיעות ביחיד אבל ניתן לשבץ אותן לפי הסדר פלא להסתכל קדימה). לכן,

$$C_k \le \frac{1}{m}\sum_{j\neq k} p_j + p_k \le \frac{1}{m}\sum_j p_j + \left(1-\frac{1}{m}\right)p_k$$

$$\le C^* + \left(1-\frac{1}{m}\right)C^*$$

$$= \left(2-\frac{1}{m}\right)C^*$$

זמן הסיום שלה שווה הזמן שיוצא לאלגוריתם

המקרה הגרוע לאלגוריתם הזה הוא רק של הרבה עבודות קטנות ואז עבודה אחת מאוד ארוכה.

אבל יש לנו שיפורים לאלגוריתם שלנו מאפשר את העבודות לארוכות מדיוק בשביל המקרה הגרוע ומה שנחמד בהם הוא שהם נותנים קירוב שלא תלוי ב- m.

נשאלת השאלה אם השיפורים האלה הם שיפורים אמיתיים גם במציאות.

אז עשו ניסוי. קודם כל נגבר פיה שכל האלגוריתמים במציאות יותר טובים מהחסמים שהגלינו להוכיח עליהם. אבל יותר מעניין הוא שדווקא האלגוריתם המקורי היה הכי טוב וקרוב מאוד לאופטימום.

למה? כי הבידוקים תלויים ב- workload והמקרה הגרוע שלהם עשו את האופטימיזציות הלא נפוץ כל כך.

- רמה מטרידות

הדרך הנאיבית היא $O(n^2)$. אבל אם נקח בחשבון את הירכיית הזיכרון אפשר לאופטמז דברים. לדוגמא בחשבון את ה- cache, התחרות על ה- bus וכו'.

רעיון 1: לחשב בחלקים - ב- tiles שנכנסים ב- cache

רעיון 2: להשתמש ב- prefetching כדי שהנתונים של

ההצלה הבאה כבר יהיו ב- cache נשגיח עליה.
אם אכן התשובה נמצאת כן ב- cache אז ה- link בין
ה- cache לזכרון פנוי ואפשר להשתמש בו.
תיקון 3: ה- cache הוא אוטומטי אז גם צריך
לבדוק שנכתבות לאז תיתבצע אחת בשניה. אבל
אפשר להתעלם אם ה- נשלח לעמקומות שונים כדי שלא
תיווצרנה התנגשויות ב- cache.

אז אימשו אות זה וזה אכן התברר להיות יותר טוב
מהספרייה האמצעית של IBM שהשקיעו בה המון!!!

- MAXIMUM FLOW.
יש גרף עם שני קודקודים מיוחדים s, t. לכל קשת
יש קיבולת $c(u,v)$. צימוד הוא פונקציה כך ש-

$$f : V \times V \to R$$

$$f(u,v) \le c(u,v)$$

$$f(u,v) = -f(v,u)$$

$$\sum_v f(u,v) = 0$$ לכל u חוץ מ- s ו- t

המטרה הוא למצוא את הצימוד המקסימלי שאפשר
שיעבור מ- s.
יש שני אלגוריתמים לפתרון הבעיה שפסיבוכיות שלהם שונה
מאוד ותלויה מאומים שונים. אינם לבדוק איזה מהם יותר
טוב באמצעות. יש גם כל מיני אופטימיזציות לגביהם שלא
ברור איך ההשפעה האמיתית שלהן. ובאמצע אם נריץ לבדוק
את השפעת כל הטיפולים של הגרף.
כדי לענות על כל השאלות האלה עושים ניסיים!
המסקנה היא כמובן שאין תשובה אמיתית- אין גרף שמתאים טוב
לכל הגרפים מול הגבלים