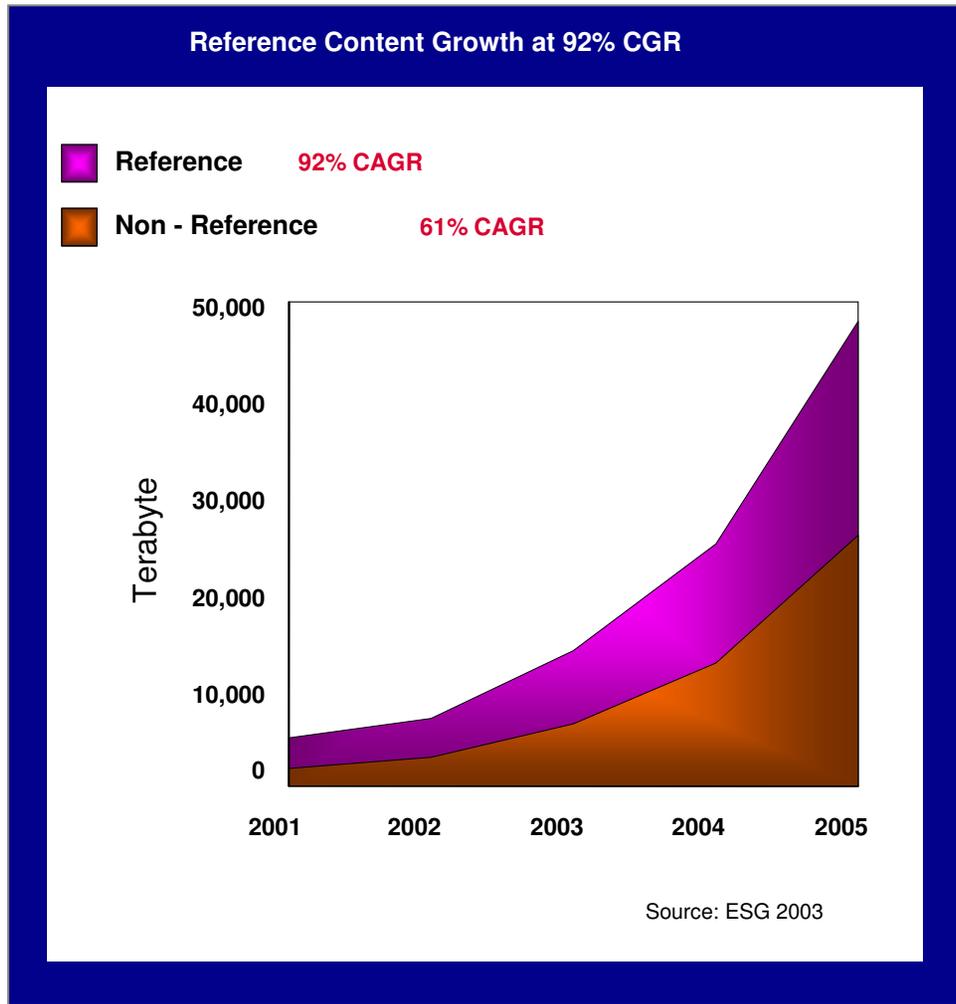


ניהול כמויות עצומות של מידע בעידן האינטרנט

דפנה שינולד

יבמ

כמות המידע הדיגיטאלי הנשמר בדיסקים הולכת וגדלה



Source: Enterprise Solutions Group 2003

נפחו של המידע הקבוע (Reference Data)

דואר אלקטרוני, מולטי מדיה, דפי תוכן

גדל בקצב מהיר יותר מאשר השאר

טרנזקציות במסדי נתונים

(שם דרושות אמינות וזמינות)

רגולציות שונות גורמות לחלק נכבד מנפח

המידע הקבוע הנשמר על ידי חברות

ואירגונים.

ביחד, מגיע הגידול ל- 75% ממוצע שנתי

משוקלל

▪ terabyte = 10^{12} bytes

▪ petabyte = 1000 terabyte = 10^{15} b

ואייך ננהל את הכמות ההולכת וגדלה הזו

וי' תר מהמה, בני הזיהר:

עשות ספרים הרבה - אין קץ, ולהג הרבה - יגעת בשר.

(קהלת יב, יב)

שלושה היבטים של האתגר

- אירגון מערכות איחסון גדולות
- אירגון ואיחזור של מידע בכמויות עצומות
- נסיון להקטין נפחים – דחיסת מידע (רק נזכיר)

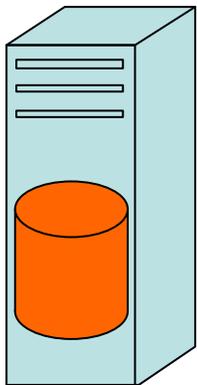
אירגון מערכות איחסון גדולות

לקראת שיתוף פעולה בין רכיבים "אינטיליגנטים"

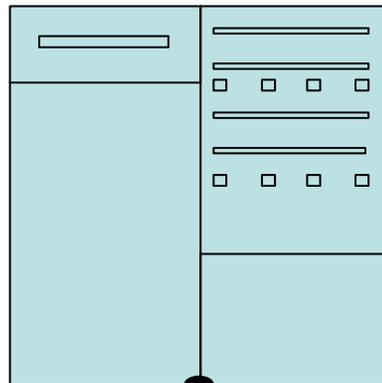
הדיסק גדל, ויוצא מהבית

from directly attached to networked attached storage

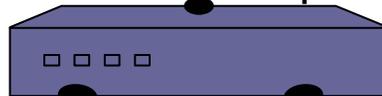
מחשב עצמאי



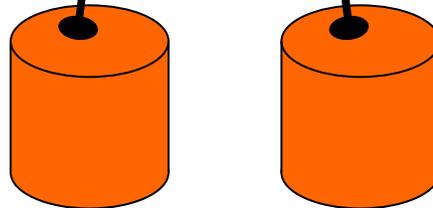
שרת יישומים – Application Server



שרת מערכת הקבצים שברשת NFS server



fiber channels



דיסקים

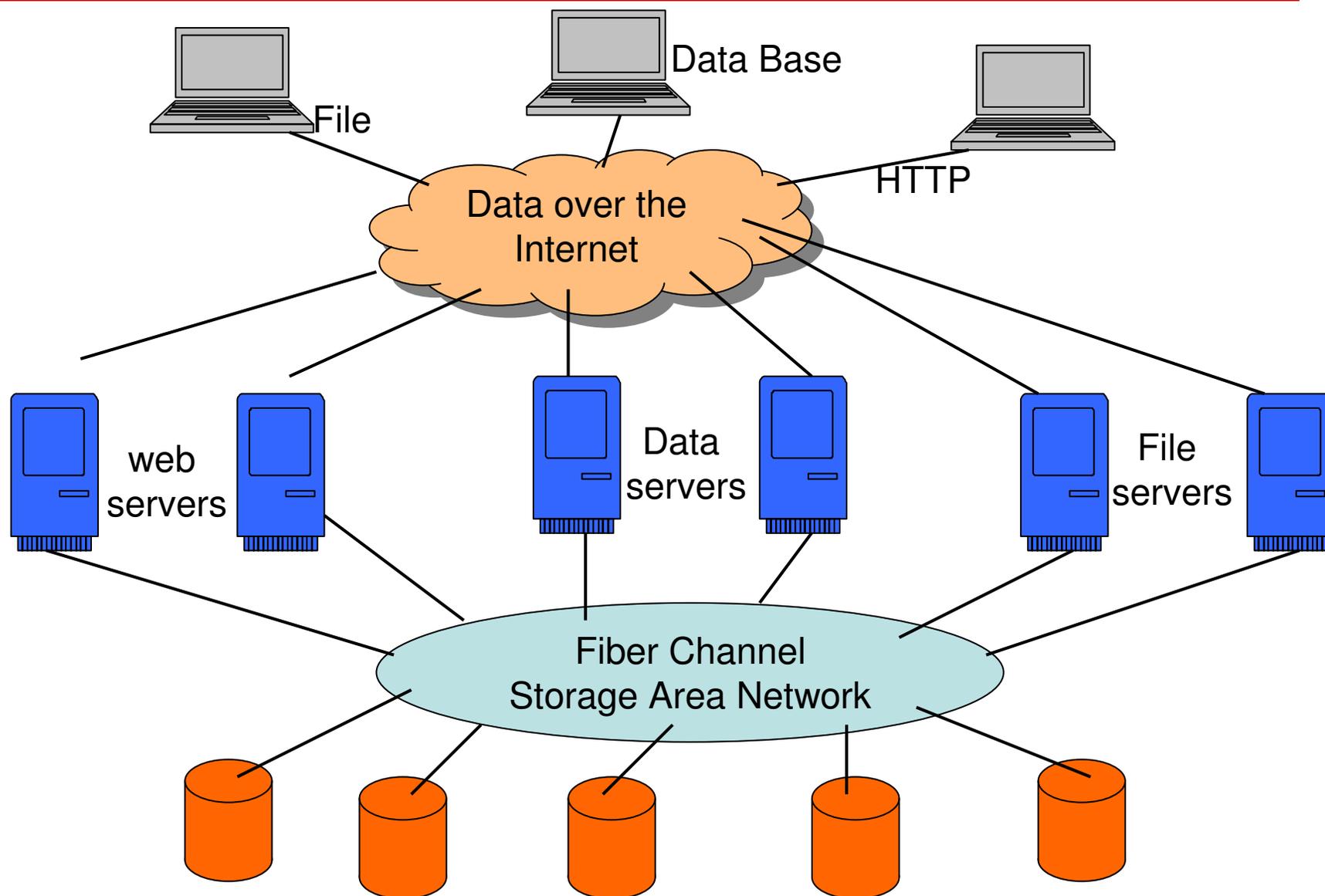
מניעים:

- מחירים יורדים של תקשורת נתונים
- שרת יעודי לנתונים
- שיתוף נתונים
- צורך בגיבויים לדיסק

Networked File System

ניצב בין הדיסקים ובין שרתי היישומים
פועל ומפעיל פרוטוקולים
לשיתוף קבצים בין היישומים השונים

והדיסק גדל עוד יותר, גדל ומתרחק



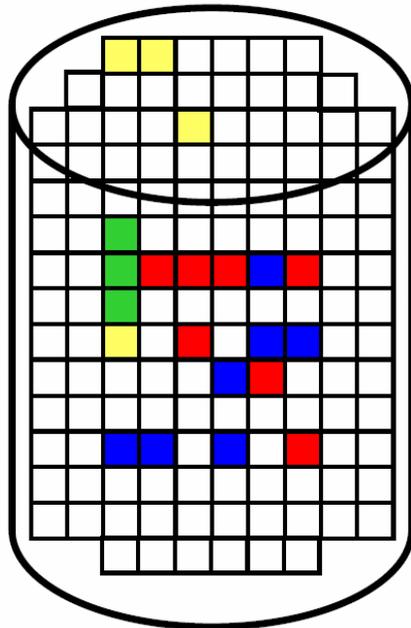
Object Store מחסן אובייקטים

הדיסק מגלה עצמאות "ומגדיל ראש"

פעולות:

קרא בלוק מס' i
כתוב תוכן חדש לבלוק מס' j

קביעת מקום בדיסק:
בקביעה חיצונית (מערכת קבצים)

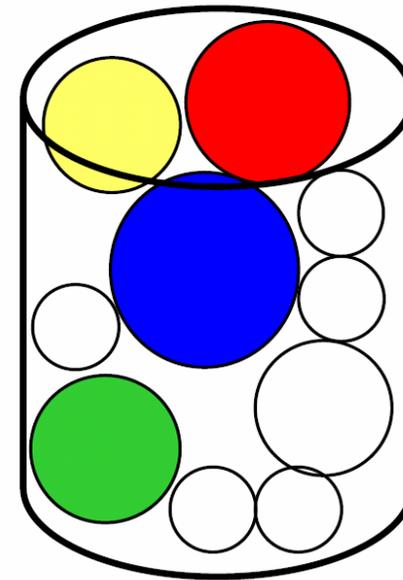


Block Store היום: דיסק של בלוקים

פעולות:

קרא אובייקט מס' i
כתוב תוכן חדש לאובייקט מס' j
צור אובייקט וקבע מספרו
מחק אובייקט מס' k

קביעת מקום בדיסק:
בניהול מקומי



מחר: מחסן אובייקטים Object Store

ה- Web הענק

איך נמצא מחט בערמה גדולה של שחת

הגודל והדינמיות של ה- Web

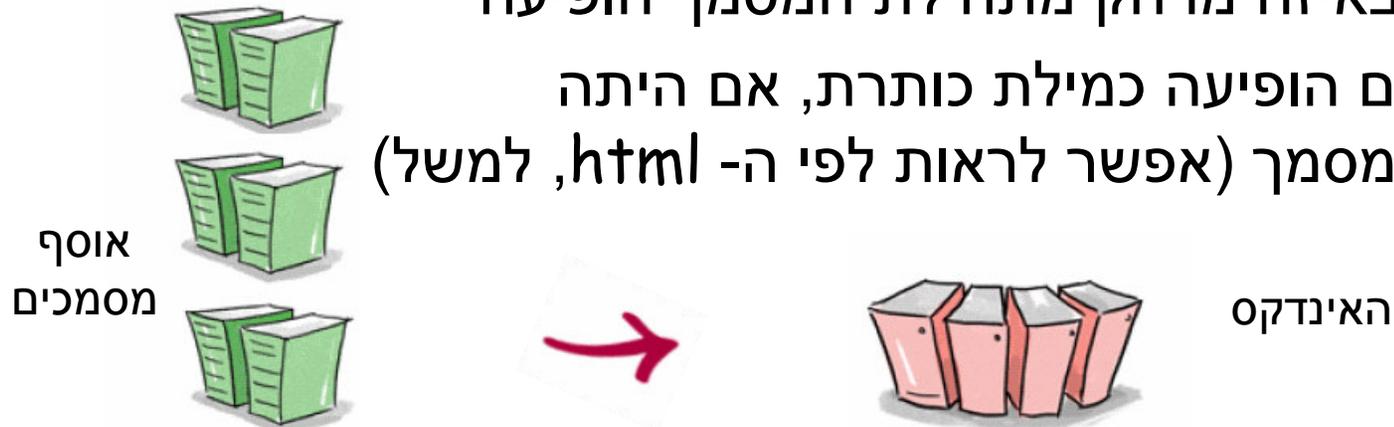
- מליארדי דפים (Google מדווחים על 8 מליארד)
- מליוני אתרים
- עשרות terabytes של מידע דיגיטאלי
- הגידול החזוי: הכפלה מדי שנתים עד שלוש
- דפים חדשים נוספים ודפים מורדים
- מידע לא מובנה – טקסט חופשי בכל צורה ואופן
- צורות שונות של קבצי המידע (html, doc, pdf, free text)
- מידע מפוזר ומבוזר על פני יבשות תבל ואירגונים שונים
- (התפשטות וגדילה כמו אחרי המפץ הגדול)

איחזור מידע Information Retrieval

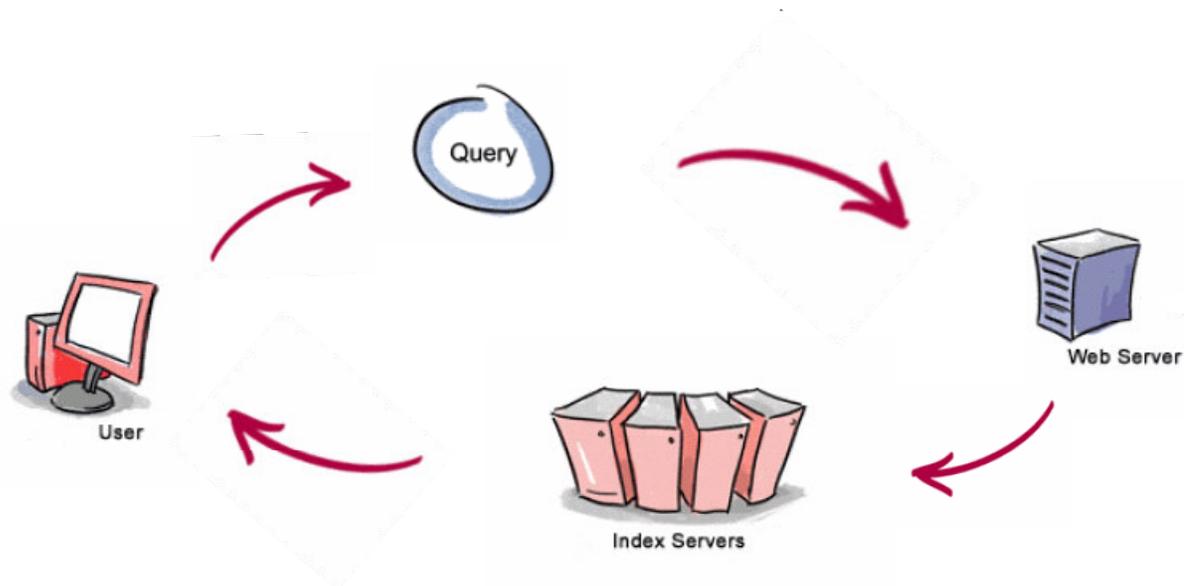
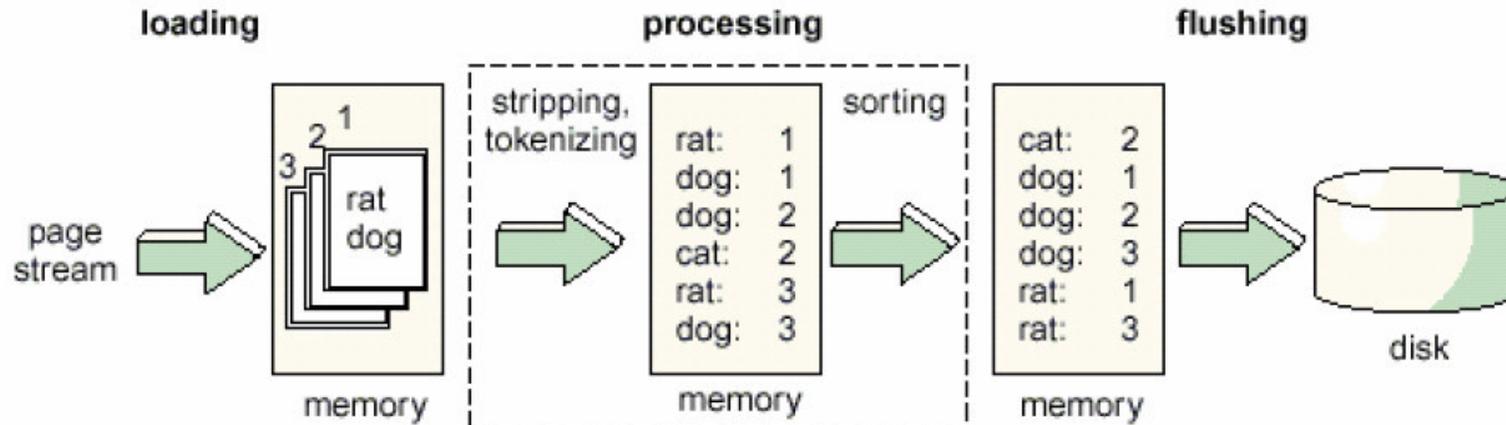
- בראשיתו, היה זה מקצוע הספרנות
- תיוג לפי נושאים (classification)
- עיבוד שפות טבעיות והבנת הנושאים (מומחיות לנושאים)
- המידע מסודר לחיפוש ולדפדוף (retrieval and browsing)
- התפתחות ה-Internet שינתה את מהות המקצוע לבלי הכר
- מאגר מידע אחד ענק לכל הנושאים
- גישה חופשית (או זולה) ופשוטה לכל
- מותר לכולם לפרסם הכל – אין וועדת עורכים, ולא עורך ראשי
- מנוע חיפוש צריך לאפשר חיפוש ושליפה, וגם לאפשר דפדוף – סקירה על פי נושאים ותתי נושאים (directories).

האינדקס הגדול

- כמו הקונקורדנציה של התנ"ך:
- בהנתן מילה, מחזיר את רשימת המופעים שלה בכל התנך
- מנוע החיפוש בונה אינדקס (קרוי גם אינדקס הפוך *inverted index*)
- משוטט (*crawl*) בכל ה- Web מדף לדף, בעקבות הקישורים, ובאופן אקראי, ואוסף את כל המסמכים ואת כל המילים שפוגש
- מכין את רשימת כל המילים השונות שראה (הלקסיקון)
- עבור כל מילה, יוצר רשימה של כל מופעיה: באיזה מסמך (מה ה- *url* שלו) ובאיזה מרחק מתחילת המסמך הופיעה
- מציין גם אם הופיעה כמילת כותרת, אם היתה מודגשת במסמך (אפשר לראות לפי ה- *html*, למשל)



האינדקס



דחיסה

מנוע החיפוש קיבל שאילתה בת שתיים שלוש מילים ומצא מהר את כל הדפים שמכילים את מילות השאילתה איך ידרג אותם?

- השאיפה היא להחזיר מסמך מספק בין עשר התוצאות הראשונות
 - (above the folder)
- הצורה בה מופיעה כל מילת שאילתה במסמך תורמת לדירוגו:
 - האם מופיעה בכותרת? האם מופיעה מודגשת?
 - האם מופיעה פעמים רבות במסמך?
- איך מספר הופעותיה במסמך מתיחס למספר הופעותיה באוסף כולו?
- המילים במסמכים אינן שכיחות במידה שווה
- ישנן המופיעות פעמים רבות באוסף המסמכים, וישנן שמעט.
- האם מילות השאילתה מופיעות במסמך במקבץ?
- מהי הקישוריות בין המסמכים (ענין חדש -- מיד בהמשך)
- האם השאילתה מבקשת מידע כללי, או נועדה להוליך לאתר מסחרי ("world war ii" "EIAI")

חוק Zipf לשכיחות ההופעות של מילים בשפה בלשן אמריקאי, 1902 - 1950

Government documents, 157734 tokens, 32259 unique

8164 the	969 on
4771 of	915 FT
4005 to	883 Mr
2834 a	860 was
2827 and	855 be
2802 in	849 Pounds
1592 The	798 TEXT
1370 for	798 PUB
1326 is	798 PROFILE
1324 s	798 PAGE
1194 that	798 HEADLINE
973 by	798 DOCNO

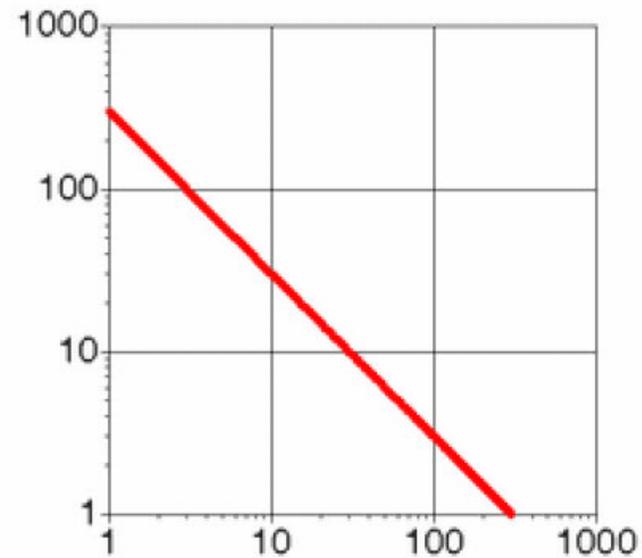
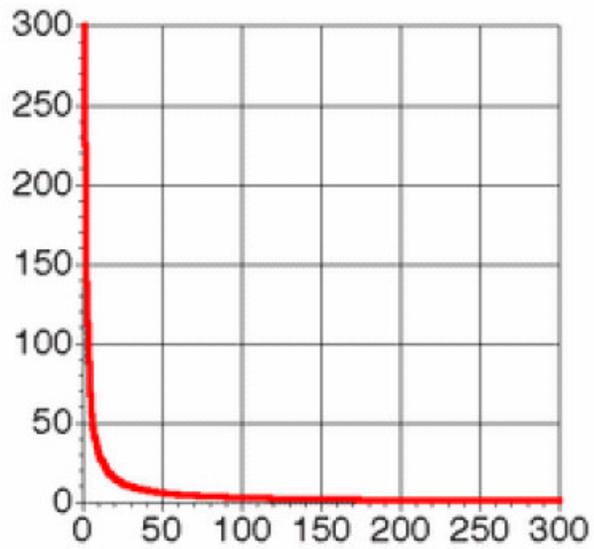
1 ABC
1 ABFT
1 ABOUT
1 ACFT
1 ACI
1 ACQUI
1 ACQUISITIONS
1 ACSIS
1 ADFT
1 ADVISERS
1 AE

השכיחות היחסית של המילה ה- n ית בדרוג המילים על פי שכיחותן פרופורציוני ל- $1/n^a$ עם a קרוב מאוד ל-1.
כלומר: עבור קבוע C , המילה השכיחה ביותר מופיעה בערך C פעמים, המילה השנייה בשכיחותה – $C/2$ המילה השלישית – $C/3$ וכו'

Power Law

התפלגות Zipf

בסקלה לינארית, וב-Log-Log

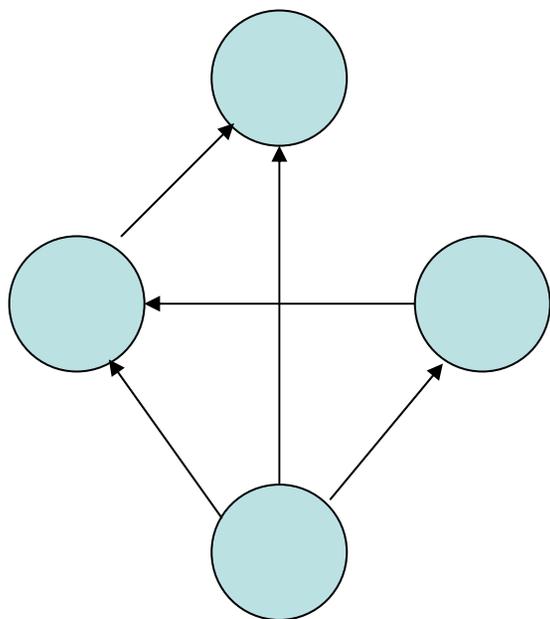


הקישוריות היתה כאן תמיד

- נולדה כפרק באיחזור מידע קלאסי -- Citation analysis
 - איזה מאמר / ספר מצטט איזה מאמר / ספר
 - איזה עיתון מצטט איזה עיתון
 - איזה כותב מצטט איזה כותב
- לאורך כמה שנים, בממוצע, מצוטט מאמר שמתפרסם בעיתון זה

ה-Web כגרף

- הדפים הם הצמתים, והקישורים (links) – קשתות מכוונות
- מהדף המצביע אל הדף המוצבע

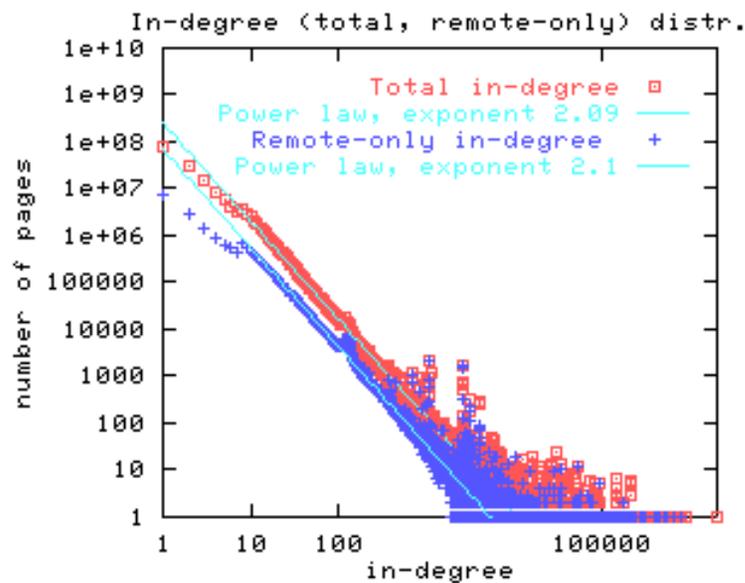


- אייך מפולגים הצמתים לפי דרגות הכניסה? דרגות היציאה?
- האם הגרף קשיר?
- קישור פנימי (בתוך אותו אתר) לעומת קישור מבחוץ

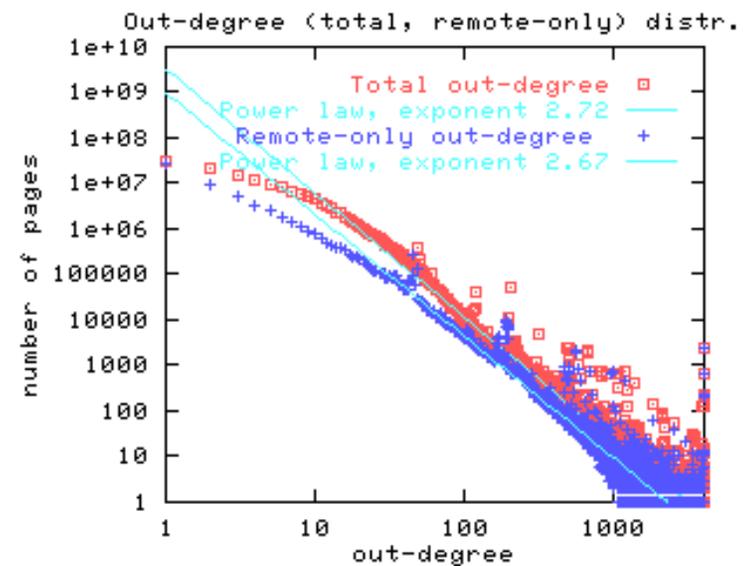
תשובות מענינות:

התפלגות הצמתים

על פי דרגת הכניסה ועל פי דרגת היציאה



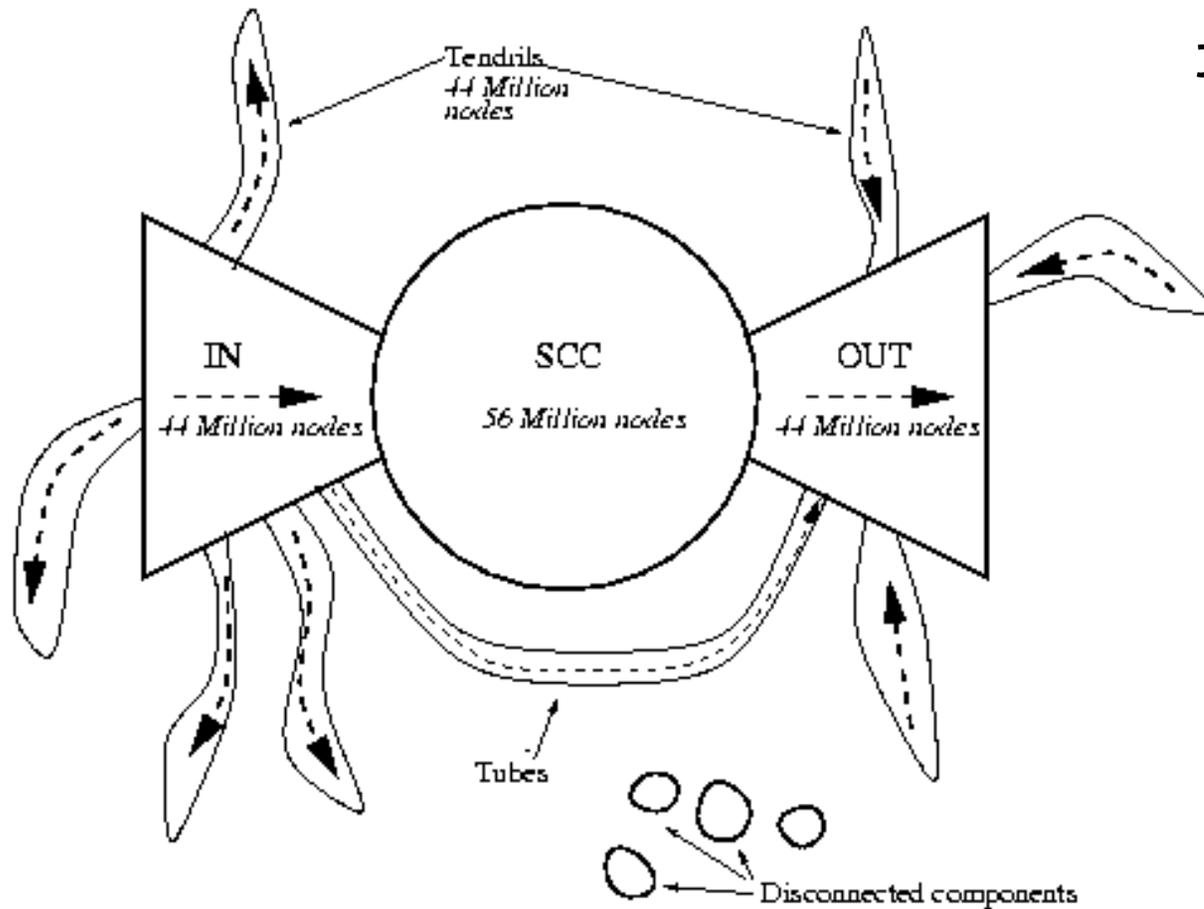
דרגת כניסה



דרגת יציאה

Power Law מופיע שוב

הקישוריות של גרף ה- Web



תופעה דומה מתגלית
גם בתתי-גרפים של
ה Web
(תופעה אנושית?)

מה אפשר ללמוד מהקישורים

- Anchor text: בתוך מסמך A, בסמוך לקישור שיוצא ממנו אל מסמך B, מופיע טקסט שמתמצת את תוכנו של B.
- מאנדקסים את מילות התמצית האלה כאילו היו מילים של B
- מיהו מסמך חשוב: מי שרבים מצביעים עליו (authority)
- מיהו מסמך שמכיל מדריכים (קישור לענינים שונים): מי שיוצאים ממנו קישורים רבים (hub)
- מיהם מסמכים שעוסקים בענינים דומים:
 - מצביעים זה על זה
 - מצביעים על אותם מסמכים אחרים
 - מוצבעים על ידי אותם מסמכים אחרים

Lawrence Page and Sergey Brin from their time as graduate students at Stanford University.

- עוד בהיותם סטודנטים, פירסמו מיסדי **Google** מאמר בענין דירוג על פי קישורים
- הדירוג נקרא **Page Rank** והוא כנראה הבסיס להצלחה.
- הדירוג בפועל היום חסוי, וודאי משתנה מדי פעם
- מבוסס על דירוג על פי קישורית נכנסת
- **אבל:** לא בפשטות המספר הכולל של קישורים נכנסים
- שחשוף ל **spammers**: אתרים שמוליכים שולל את המנוע
- **אלא:** קישורים ממסמכים חשובים משוקללים יותר מקישורים ממסמכים שאינם חשובים.
- הדירוג של מסמך נקבע על פי הדירוג של המסמכים המכילים קישור אליו

PageRank

$$PR(A) = \frac{(1 - d)}{N} + d \left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \dots + \frac{PR(T_n)}{C(T_n)} \right)$$

■ ופה:

- $PR(A)$ הוא הדרוג Page של A
- $PR(T_i)$ הוא הדרוג Page של T_i
- $C(T_i)$ הוא המספר הכולל של קישורים שיוצאים מ- T_i
- d גורם בין 0 ל-1 (לדברי המאמר – בסביבות 0.85)
- N הוא המספר הכולל של הדפים באוסף

הסבר אינטואיטיבי הגולש האקראי

- הגולש מדף לדף בוחר בצורה אקראית אחידה, על פני הקישורים היוצאים מהדף שמבקר בו כעת, את הדף הבא שיבקר בו
- לפעמים, בהסתברות $1-d$, נמאס לגולש שלנו לעקוב אחרי הקישורים, והוא נוחת על דף כלשהו בצורה אקראית.
- בשכיחות פרופורציונית ל- $PR(A)$ יבקר גולש זה בדף A .

ואיך מחשבים את כל זה

- בנוסף לכל המילים (האינדקס) יש לשמור עכשיו על כל הקישורים ולשלוף בקלות – צריך דחיסה טובה
- מתחילים בדרוג ראשוני, ועורכים כמה איטרציות עד להתכנסות מספקת.

התפתחות יכולת החיפוש

- בענין נפח המידע המצוי אצל המנוע:
- מאמצי דחיסה של כל אוסף (קורפוס) עם כל הקישורים שעליו
- חלקים שונים ראויים לדחיסה שונה
- בענין דרוג:
- בתחילה – שיטות קלאסיות של איחזור מידע
- מסוף שנות ה - 90 -- דגש על אנליזה של קישוריות
- באופנה עכשיו – התאמות אישיות, או לפחות גיאוגראפיות
- בעתיד – Semantic Search : המנוע יבין מושגים מתוך המסמך ובאיזה ענינים הוא דן, גם מבלי שאלו יצוינו במפורש

ספרים הרבה – אין קץ
(וגם להג הרבה)
אך לא יגיעת בשר בלבד