גירסה 1.00 – 28.3.2005

מנוע החיפוש גוגל - סקירה

ניר אדר

Nir Adar

Email: underwar@hotmail.com

Home Page: http://underwar.livedns.co.il

אנא שלחו תיקונים והערות אל המחבר.

מנוע החיפוש גוגל – סקירה

גוגל – רקע היסטורי

מנוע גוגל הוקם על ידי שני סטודנטים מאוניברסיטת סטנפורד, Larry Page ו-Sergey Brin. הרעיון ההתחלתי סביבו הם פעלו היה מנוע חיפוש המתמקד בניתוח הקישורים בין הדפים ברשת האינטרנט. כמו כן מטרתם היתה ליצור מנוע שיוכל לרוץ על מחשב פשוט, ולא על מחשבי על. השם שניתן למנוע גוגל (Google) הוא משחק מילים מסביב מספר גוגול – 1 ולאחריו 100 אפסים.

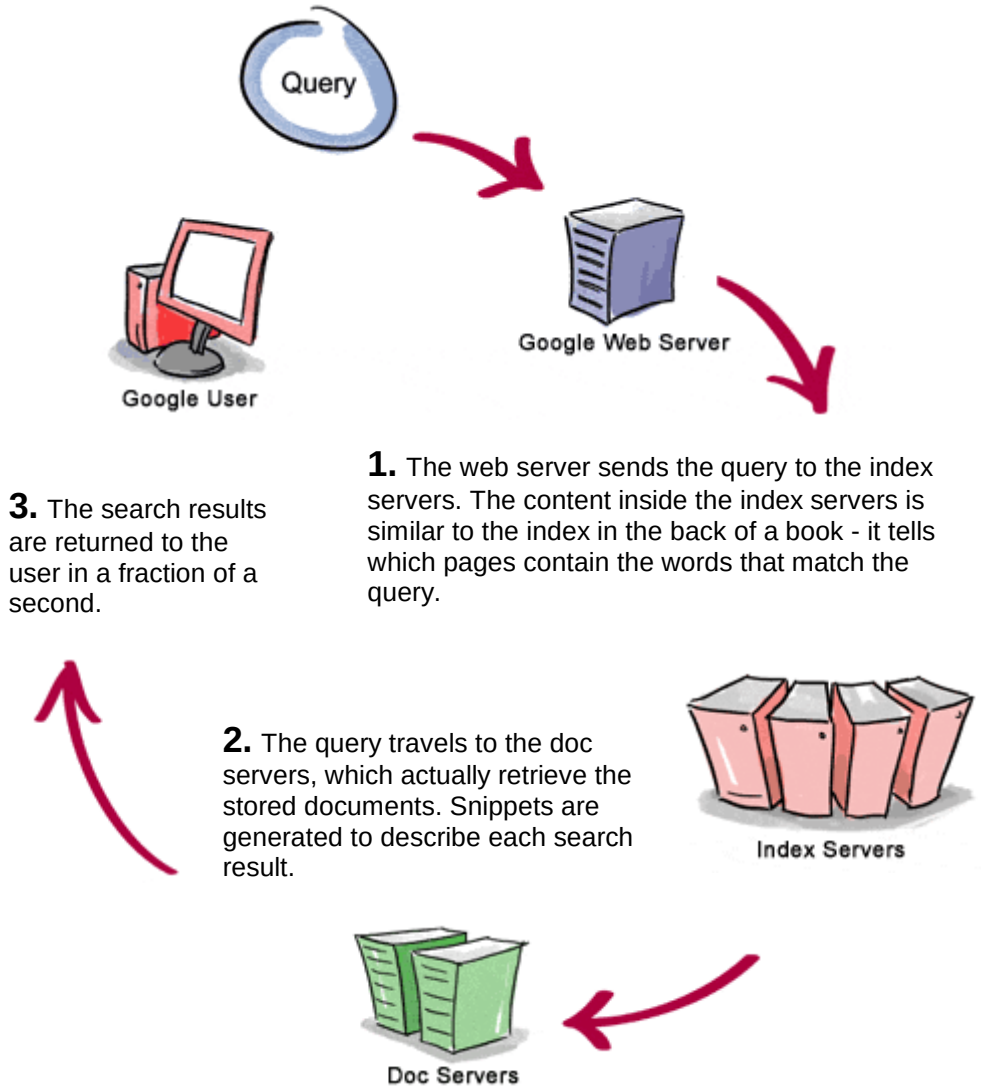
גוגל נפתח לראשונה לקהל הרחב בשנת 1998. בשנה זו כבר היו קיימים מנועי החיפוש כגון Yahoo ו-Altavista ששלטו באותה תקופה בשוק. אם זאת, תוך זמן לא רב גוגל הצליח להפוך להיות מנוע החיפוש הפופולרי ביותר. הסיבות העיקריות להצלחתו הן הרלוונטיות הרבה של התוצאות שבו, מיעוט הקישורים השבורים בתוצאות, והעובדה שבעלי אתרים אינם יכולים לקנות מקום בתוצאות החיפוש, ולכן התוצאות נחשבות אמינות ולא משוחדות.

כיצד מתבצעת שאילתה במנוע החיפוש?

Google מורכב כיום מעשרות אלפי מחשבים הפועלים יחדיו.
באופן פשטני ניתן לחלק את מערך המחשבים שלהם למספר סוגי שרתים:

- שרתי Web (Web Servers), שהם השרתים המציגים לגולשים את ממשק Google ומולם מתבצעת הפעולה של הלקוחות.
- שרתי אינדקס (Index Servers) המסוגלים לענות על השאלה "אילו מהדפים ברשת האינטרנט מכילים את המידע המתאים עבור השאילתה שהקיש הגולש".
- שרתי מסמכים (Doc Servers) המכילים עותק של הדפים. מהם Google יוצרים את הדף המוחזר אל הגולש, הכולל קטעים מן האתרים המבוקשים. למעשה ניתן להגיד גוגל מייצרים אצלם העתק של רשת האינטרנט (לפחות חלק מרשת האינטרנט – החלק אותו ניתן לחפש בגוגל).

השרטוט הבא לקוח מאתר Google. השרטוט מתאר את התהליך המתבצע כאשר משתמש שולח שאילתה אל Google. השרטוט נלקח מהכתובת http://www.google.com/corporate/tech.html.



נסביר:

1. המשתמש שולח שאילתה אל שרת Web של Google.
2. שרת ה-Web פונה לשרתי האינדקס שימצאו את הדפים המכילים את המידע.
3. השאילתה עוברת אל שרתי המסמכים, בהם מורכב דף התוצאה.
4. דף התוצאה מוחזר אל המשתמש.

כיצד גוגל מדרג את תוצאות החיפוש?

כאשר אנחנו מחפשים מילה כלשהי, ייתכן שיהיו אלפים ואף מאות אלפים של דפים הכוללים מילה זו. שאלת השאלות סביבה מסתובבת כל תעשיית קידום האתרים היא – כיצד גוגל מחליט איזה דף חשוב יותר מאחר, ויש להציבו גבוה בראש תוצאות החיפוש.

Pagerank

Google קובע את מיקום הדף על ידי מספר מדדים. אחד מהם, שהיה פעם העיקרי וכיום חשיבותו ירדה, הינו אלגוריתם Pagerank.

PageRank זהו מספר בו משתמש Google כדי להגדיר את מידת החשיבות של דף מסויים באינטרנט. מדד זה נקבע על פי מספר האתרים המקשרים אל האתר. **הרעיון המנחה** - ככל שיש יותר קישורים נכנסים אל דף מסויים ברשת, כך הדף כנראה חשוב יותר. בעצם – כל אתר המקשר אל דף מסוים, "מצביע אמון" בדף זה.

אלגוריתם Hilltop

אלגוריתם נוסף הקובע רבות בדירוג האתרים של Google הוא אלגוריתם Hilltop, הפועל בנוסף לאלגוריתם ה-PageRank.

הרעיון שעומד מאחוריו הוא קיומם של "דפים מוסמכים" בנושאים שונים. למשל דפים העוסקים בנושא מחשבים, דפים העוסקים בנושא דייג וכו'. לפי אלגוריתם זה, במידה ונרצה לבנות אתר המציג דגמי סירות, ייתכן שנעדיף קישור מאתר של דייגים מאשר מאתר בנושא בניית אתרים. בנוסף – גם הטקסט של הקישור משנה – אם נקשר אל האתר שלנו עם מילת מפתח מסויימת, הדבר יעלה את דרוג האתר שלנו עבור מילה זו.

מבנה הדף

גורם חשוב נוסף בדירוג אתר הינו מבנה הדף – אם האתר מכיל את המילים אותן מחפש הגולש, ונראה לגוגל לפי מבנה הדף כי אכן נושא הדף הוא מילים אלו, לדף יש סיכוי להגיע למקום טוב בתוצאות החיפוש.

פרמטרים נוספים

גוגל משתמשים בפרמטרים רבים נוספים על מנת לקבוע את מיקומו המדויק של האתר. לטענת גוגל, מרכיבי החישוב כוללים מאות פרמטרים שונים.

עדכון התוצאות בגוגל

מרגע שאתר כלשהו התווסף לאינטרנט, או מרגע שנוסף/השתנה דף חדש לאתר – תוך כמה זמן תוצאות החיפוש בגוגל מתעדכנות עם האתר החדש?

התשובה תלויה בחשיבות שגוגל נותן לאתר וכן במבנה האתר. אתרים להם חשיבות גדולה – גוגל יבקר בהם פעמים רבות ויסרוק אותם לשינויים. גוגל נותן לרוב חשיבות לאתרים עם PageRank גבוה, וכן לאתרים המתעדכנים לעתים קרובות. ייתכן אפילו מצב שגוגל יבקר באותו אתר עשרות פעמים ביום על מנת לאתר עדכונים. אתרים בעלי PageRank נמוך, או אתרים שלא התעדכנו תקופה ארוכה, יזכו לפחות ביקורים מרובוט החיפוש של גוגל.
מהרגע שרובוט החיפוש סורק אתר עד לרגע בו תוצאות החיפוש של גוגל מתעדכנות – עוברת תקופה של כ-24-48 שעות לפחות ("תקופת צינון") עד שהנתונים מתעדכנים בתוצאות החיפוש.
זמן העדכון המדויק – אפילו עבור אותו אתר, משתנה כל הזמן בתקופה האחרונה. Google משנים כל העת את האלגוריתם שלהם, את קצב העדכון וכדו' – גם לצורך שיפור התוצאות וגם על מנת להקשות על גורמים שונים להבין בדיוק כיצד המנוע עובד, ולגרום למניפולציות רציניות בתוצאות החיפוש.

המבנה של גוגל, Google Datacenters

Datacenters – חלוקת עומס הגושלים, מספר מרכזים המכילים את המידע

כיצד גוגל עונה לשאלות החיפוש של מליוני האנשים הפונים אליו בכל יום? לא הגיוני שכל האנשים מגיעים אל אותו שרת, ומקבלים ממנו שרותים. על מנת לתת לפזר את עומס הפניות של גוגל, וגם על מנת להבטיח שאם חלק מהמחשבים קורסים מנוע החיפוש ימשיך לעבוד ללא הפסקה, ל-Google אין מרכז ראשי אחד אלא מספר מרכזי מידע שונים. מנוע החיפוש Google מורכב מ-100,000 שרתים המחולקים לקבוצות המכונות datacenters.
כאשר אנשים פונים לכתובת google.com, הם מופנים כל העת אל שרתים שונים שיטפלו בפניות שלהם, על מנת לפזר את העומס ביניהם.
בייחוד ברגעים בהם גוגל מעדכן את בסיסי הנתונים שלו באתרים חדשים ובשינויים באתרים קיימים, datacenters שונים מציגים תוצאות שונות עבור חיפושים.

מקדמי אתרים עוקבים לרוב אחרי datacenters שונים כאשר הם מקדמים אתרים, על מנת לראות את המגמה של תוצאות הקידום – לרוב מספר datacenters מתעדכנים לפני אחרים, וכך ניתן לראות את השפעת תהליך הקידום – לפני שכל השרתים התעדכנו.

החומרה בה משתמש גוגל

החומרה בה משתמש גוגל היא לרוב שרתי Intel 2U בעלי מעבד Xeon עם ארכיטקטורה הדומה לזו של מחשב אישי סטנדרטי, וכן הם משתמשים בכונני IDE. המידע בשרתים נשמר על מערכת קבצים בשם GFS - Google File System. מערכת זו נכתבה במיוחד על ידי הצוות של גוגל על מנת להתאים לעבודה של מנוע החיפוש.