Εισαγωγή στην πληροφορική

Ενότητα 11 : **Ο αλγόριθμος PageRank της Google**

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

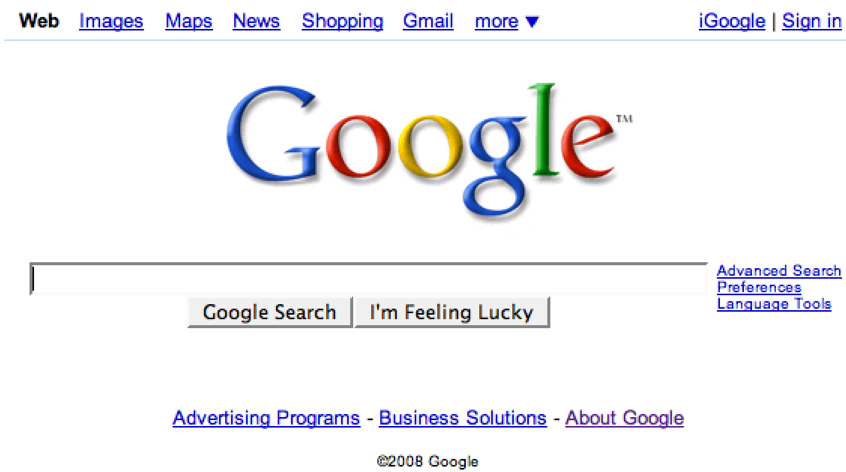
Χρηματοδότηση

- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «**Εκπαίδευση και Δια Βίου Μάθηση**» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο ΤΕΙ Ηπείρου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Η μηχανή αναζήτησης Google

- Το Google ξεκίνησε σαν μια κολεγιακή εργασία από τον Larry Page και τον Sergey Brin το 1996 με σκοπό την κατασκευή μιας μηχανής αναζήτησης
- Το όνομα Google είναι παραφθορά του Googol = 10^{100} (1 μονάδα ακολουθούμενη από 100 μηδενικά)
- Το όνομα της εταιρείας έχει γίνει ρήμα (Google it!)
- Αριθμός υπαλλήλων: 28768 (2011-06-30)



10.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000

Work in Progress

- Κάθε χρόνο η Google πραγματοποιεί περί τις 500 βελτιώσεις στην μηχανή αναζήτησής της.
- Κάθε 2 περίπου χρόνια η Google κάνει μια σημαντική αναβάθμιση στην μηχανή αναζήτησης της.
- Το γεγονός ότι η αναβάθμιση γίνεται ενώ η μηχανή αναζήτησης είναι σε λειτουργία αποτελεί μεγάλη τεχνολογική πρόκληση.

Google’s Mission Statement

Οργάνωση της πληροφορίας όλου του κόσμου.

Google’s slogan: Don’t be evil.

Ανταγωνιστές

- Yahoo!
- **Microsoft Bing:** Δυναμικές εξαγορές (π.χ. FlatRate) και συμφωνίες (Yahoo). Εστιάζει σε κάθετες αγορές (π.χ. αεροπορικά εισιτήρια, υγεία)
- Ask
- HotBot
- Altavista

Ιστορία της μηχανής αναζήτησης Google (1997-…)

- Σεπτέμβριος 1997. Η μηχανή αναζήτησης ονομάζεται Google από Backrub. Κατατάσσει τις σελίδες με βάση τον αριθμό και την ποιότητα των εισερχόμενων συνδέσμων
- Αύγουστος 2001. Ο αλγόριθμος γράφετε από την αρχή για να είναι εύκολη η προσθήκη νέων κριτηρίων
- Φεβρουάριος 2003. Η Google λαμβάνει την πρώτη της πατέντα για την «τοπική ανάλυση συνδεσιμότητας» (local connectivity analysis) σύμφωνα με την οποία δίνεται μεγαλύτερη βαρύτητα σε σημαντικές ιστοσελίδες.
- Καλοκαίρι 2003. Fritz. Ο κώδικας αναβαθμίζεται έτσι ώστε να ενημερώνει τους καταλόγους συνεχώς και όχι σε δέσμες.
- Ιούνιος 2005. Διάθεση αποτελεσμάτων που εξαρτώνται από προηγούμενες αναζητήσεις του χρήστη.
- Δεκέμβριος 2005. Αναβάθμιση κώδικα έτσι ώστε να γίνεται αναλυτικότερη αναζήτηση για πληροφορίες που μπορούν να συλλεχθούν από το διαδίκτυο.
- Μάιος 2007. Αναζήτηση με ταυτόχρονη παρουσίαση αποτελεσμάτων σε ιστοσελίδες, εικόνες, βίντεο, ειδήσεις και βιβλία
- Δεκέμβριος 2009. Αναζήτηση σε πραγματικό χρόνο. Εμφάνιση σελίδων από blogs και Twitter μόλις ανέβουν στο Internet

Web search

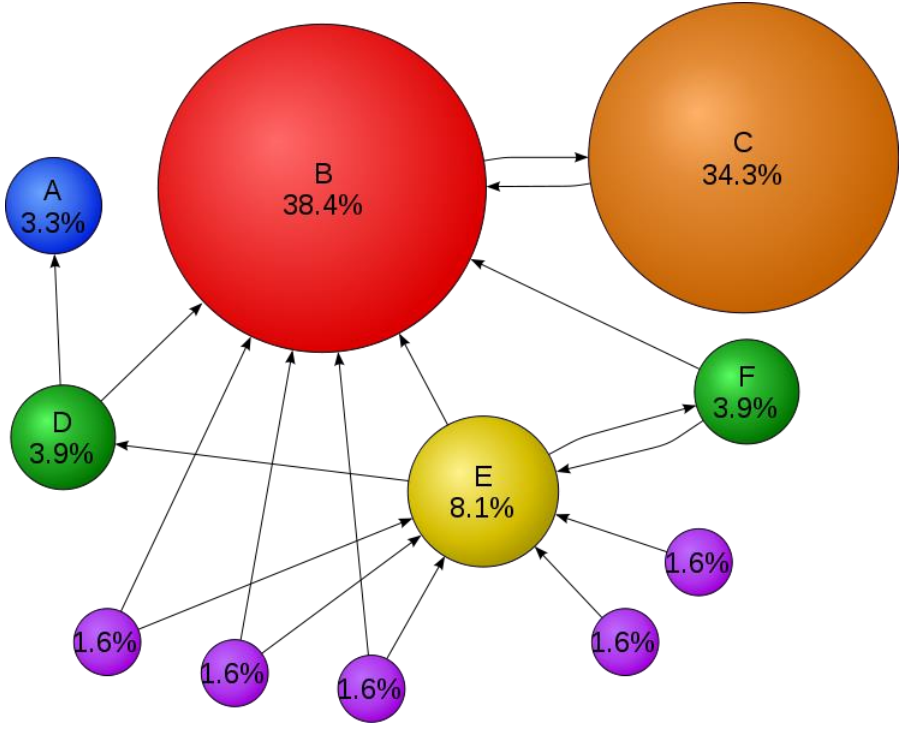
1. Προγράμματα της **Google (spiders)** περιφέρονται στον παγκόσμιο ιστό και συλλέγουν τα περιεχόμενα όλων των σελίδων που είναι προσπελάσιμες.
2. Τα δεδομένα αυτά ταξινομούνται σε ένα ευρετήριο οργανωμένο κατά λέξη.
3. Κάθε φορά που ο χρήστης κάνει μια αναζήτηση εντοπίζονται στο ευρετήριο σχετικές σελίδες δημιουργώντας μια λίστα με μέγεθος εκατοντάδων, χιλιάδων ή και εκατομμυρίων σελίδων.
4. Το απαιτητικότερο έργο είναι η κατάταξη των αποτελεσμάτων έτσι ώστε να αποφασιστεί ποια θα είναι τα αποτελέσματα που θα φαίνονται πρώτα στην λίστα.

Αλγόριθμος PageRank

- Αν και η Google έχει δημοσιοποιήσει κάποια γενικά στοιχεία για τον αλγόριθμο οι λεπτομέρειες υλοποίησης του είναι εταιρικό μυστικό.
- Ο αλγόριθμος αναθέτει σε κάθε σελίδα που επιστρέφεται ως αποτέλεσμα μια βαθμολογία συνάφειας (relevancy score) που την ονομάζει pagerank και η οποία εξαρτάται:
  - Την συχνότητα και την θέση των λέξεων κλειδιών στην ιστοσελίδα. Αν οι λέξεις κλειδιά εμφανίζονται λίγες φορές λαμβάνει χαμηλό βαθμό.
  - Πόσο χρόνο υπάρχει η σελίδα. Βαθμολογούνται υψηλότερα οι σελίδες με καθιερωμένη παρουσία στο Internet.
  - Τον αριθμό από άλλες σελίδες που δείχνουν προς την σελίδα υπό εξέταση. Μεγάλος αριθμός τέτοιων σελίδων ανεβάζει την βαθμολογία της σελίδας.
  - Εκτός από τους παραπάνω παράγοντες υπάρχουν περίπου 200 άλλοι παράγοντες που συνδιαμορφώνουν την βαθμολογία της κάθε σελίδας.

Συλλογική Νοημοσύνη του Pagerank

- Η σελίδα C έχει μεγαλύτερο pagerank από την σελίδα E παρά το ότι έχει λιγότερες συνδέσεις προς αυτήν.
- Ο μοναδικός σύνδεσμος που έχει η σελίδα C προέρχεται από μια δημοφιλή ιστοσελίδα και αυτό της δίνει μεγαλύτερη βαθμολογία.



Υπολογισμοί Pagerank

Για κάθε κόμβο σχηματίζεται η εξίσωση:

PR(A) = (1-d) + d*(PR(t1)/C(t1) + ... + PR(tn)/C(tn))

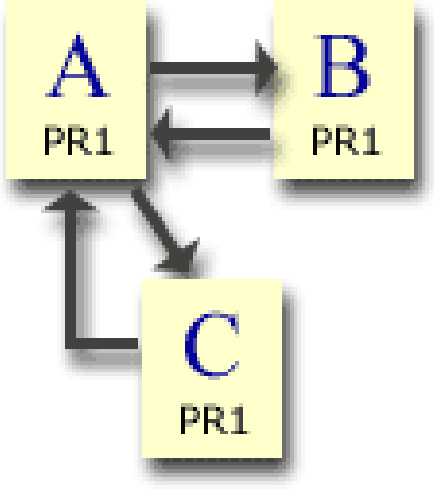
όπου t1 έως tn είναι οι κόμβοι που δείχνουν προς τον κόμβο A και C(ti) είναι ο αριθμός των συνδέσμων προς τα έξω που έχει ο κόμβος ti.

d είναι ένας συντελεστής με τιμή 0,85

Δημιουργείται ένα σύστημα εξισώσεων. Η λύση του δίνει το pagerank κάθε κόμβου σε σχέση με τους υπόλοιπους κόμβους.

Απλό παράδειγμα Pagerank

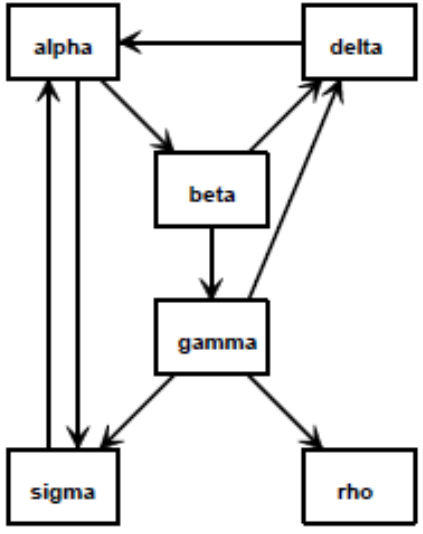
a = 0,15 + 0,85 (c + b)
b = 0,15 + 0,85 (a/2)
c = 0,15 + 0,85 (a/2)



a	1,4594	48,65%
b	0,7702	25,68%
c	0,7702	25,68%
SUM	2,9998	

Παράδειγμα Pagerank

	Είσοδοι	Έξοδοι
Alpha	2	2
Beta	1	2
Gamma	1	3
Delta	2	1
Sigma	2	1
Rho	1	0



A = 0,15 + 0,85 (S + D)
B = 0,15 + 0,85 (A/2)
G = 0,15 + 0,85 (B/2)
D = 0,15 + 0,85 (B/2 + G/3)
S = 0,15 + 0,85 (A/2 + G/3)
R = 0,15 + 0,85 (G/3)

A	1,411	32,09%
B	0,7499	17,06%
G	0,4687	10,66%
D	0,6015	13,68%
S	0,8827	20,08%
R	0,2828	6,43%
SUM	4,3966	

Πηγές

- http://www.wired.com/magazine/2010/02/ff_google_algorithm/all/1
- http://en.wikipedia.org/wiki/Google
- http://computer.howstuffworks.com/internet/basics/google.htm
- http://en.wikipedia.org/wiki/PageRank
- http://www.webworkshop.net/pagerank.html
- http://www.sirgroane.net/google-page-rank/

Βιβλιογραφία

1. Forouzan B., Mosharaf F. Εισαγωγή στην επιστήμη των υπολογιστών. Εκδόσεις Κλειδάριθμος (2010)
2. Καρολίδης Δ., Ξαρχάκος Κ.. Εισαγωγή στην πληροφορική και στο διαδίκτυο. Εκδόσεις Άβακας (2008).
3. Σφακιανάκης Μ. Εισαγωγή στην πληροφορική σκέψη. Εκδόσεις Κλειδάριθμος (2003).
4. Τσιτμηδέλης Σ., Τικτοπούλου Ε. Εισαγωγή στην πληροφορική. Πανεπιστημιακές εκδόσεις Αράκυνθος (2009).
5. Γιαγλής Γ. Εισαγωγή στην πληροφορική. Γκιούρδας εκδοτική (2009).
6. Αβούρης Ν., Κουφοπαύλου Ο., Σερπάνος Δ. Εισαγωγή στους υπολογιστές. Εκδόσεις typorama (2004).
7. Biermann A. Σπουδαίες ιδέες στην επιστήμη των υπολογιστών. Πανεπιστημιακές εκδόσεις Κρήτης (2008).
8. Brookshear J.G. Η επιστήμη των υπολογιστών, μια ολοκληρωμένη παρουσίαση. Εκδόσεις Κλειδάριθμος (2009).
9. Ceruzzi P.E. Ιστορία της υπολογιστικής τεχνολογίας. Από τον ENIAC μέχρι το διαδίκτυο. Εκδόσεις Κάτοπτρο (2006).

Σημείωμα Αναφοράς

Έκδοση: 1.0 Άρτα, 2015. Διαθέσιμο από τη δικτυακή διεύθυνση:
http://eclass.teiep.gr/OpenClass/courses/ACC136/

Σημείωμα Αδειοδότησης

[1] http://creativecommons.org/licenses/by-nc-nd/4.0/deed.el

Τέλος Ενότητας

Επεξεργασία: Α. Αναγνωστάκης

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Τέλος Ενότητας

Ο αλγόριθμος PageRank της Google