Social Web: lesson #4

- looking for relevant information
  - browsing
  - **searching**
  - monitoring
  - recommendations
- Information Retrieval
  - the inverted index
- Google.com
  - the pagerank algorithm
  - the value of words
  - the price of words

Information Retrieval

Η Αναζήτηση Πληροφοριών (ΑΠ) (Information Retrieval (IR)) είναι η επιστήμη που ασχολείται με την αναζήτηση κειμένων, ή και άλλων πληροφοριακών αντικειμένων (εικόνες, κλιπ ήχου).

Στόχος της ΑΠ είναι η αντιμετώπιση του φαινομένου του "Πληροφοριακού Υπερφόρτου", με την ανεύρεση πληροφορίας που είναι σχετική με τις παρούσες ανάγκες ενός χρήστη.

Η ΑΠ είναι ένας διεπιστημονικός τομέας που βασίζεται στην επιστήμη των υπολογιστών, τα μαθηματικά, την βιβλιοθηκονομία, την θεωρία πληροφοριάς, την ψυχολογία, την γλωσσολογία, την στατιστική και την φυσική.

Scene from Terry Gilliam's "Brazil"

inverted index

Το inverted index είναι μία δομή δεδομένων που χρησιμοποιείται ως ένα ευρετήριο που αντιστοιχεί περιεχόμενο, όπως λέξεις ή αριθμούς, στις θέσεις του σε ένα σύνολο από κείμενα ή σε μία βάση δεδομένων. Γίνεται συνήθως ο διαχωρισμός ανάμεσα record level και word level.

word \ url	url 1	url 2	...	url n
word 1	10	15	...	35
word 2	2	3	...	5
...	...	...	...	...
word n	3	6	...	8

The Inverted Index

google.com

Ξεκίνησε το 1996 ως ερευνητικό πρόγραμμα του Larry Page και του Sergey Brin, δύο διδακτορικών φοιτητών του Πανεπιστημίου του Στανφορντ. Σε αντίθεση με τις μέχρι τότε μηχανές αναζήτης που ταξινομούσαν τις ιστοσελίδες με βάση το πόσες φορές εμφανίζεται σε αυτές κάθε όρος αναζήτησης, οι Page και Brin βασίστηκαν στην υπόθεση ότι η ανάλυση των διασυνδέσεων μεταξύ ιστοσελίδων θα μπορούσε να δώσει βελτιωμένα αποτελέσματα για να δημιουργήσουν τον γνωστό πλεόν αλγόριθμο PageRank.

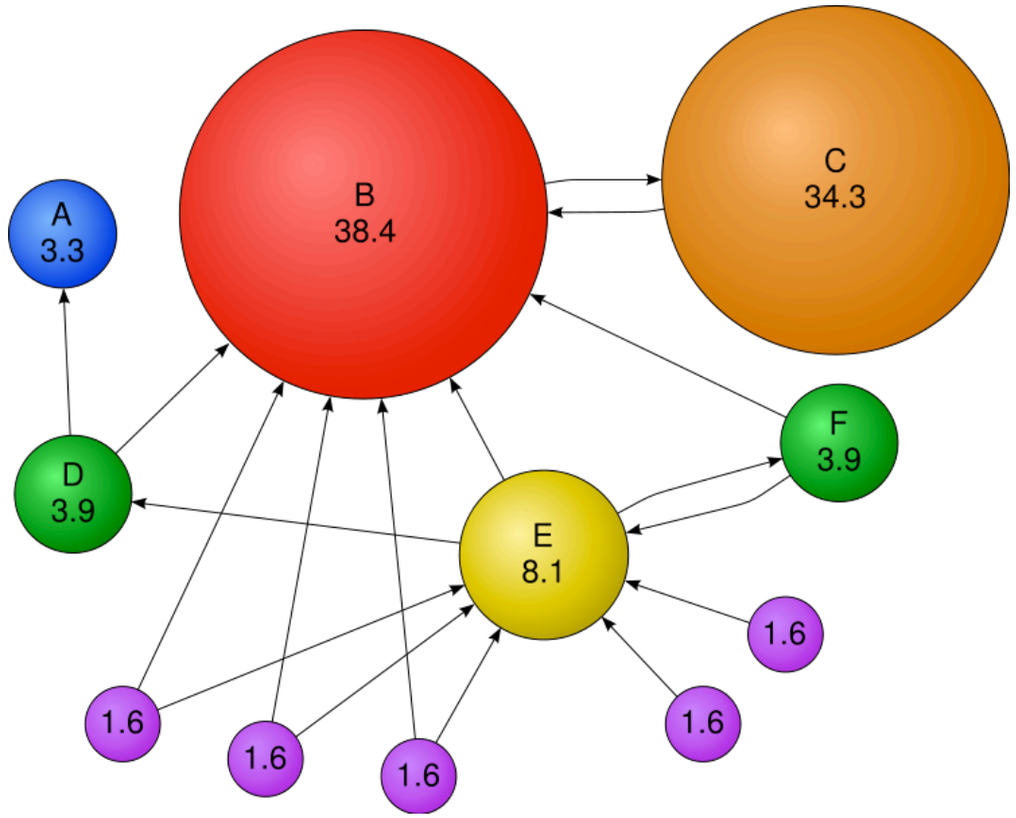
The Inverted Index

PageRank

Το PageRank είναι ένας αλγόριθμος ανάλυσης των δεσμών ενός δικτύου που προσδίδει ένα αριθμητικό βάρος σε κάθε κόμβο του δικτύου για να μετρήσει την σχετική του σημασία.

Σύμφωνα με την Google:

"Ο PageRank βασίζεται στην δημοκρατική φύση του διαδικτύου εκμεταλευόμενος την απέραντη συνδεσμολογία του ως ένα δείκτη της αξίας μίας σελίδας. Η Google αντιμετωπίζει κάθε δεσμό από την σελίδα Α στην σελίδα Β σαν μια ψήφο. Αναλύει όχι μόνο τις ψήφους που δέχεται κάθε σελίδα αλλά και το "ποιόν" των σελίδων από τις οποίες αυτές προέρχονται. Η ψήφοι που προέρχονται από σελίδες που είναι σημαντικές έχουν μεγαλύτερο ειδικό βάρος."



Η Αξία των ιστοχώρων ως συνάρτηση των συνδέσμων προς αυτούς.

the value of Words

Οι λέξεις των κειμένων έχουν στατιστική αξία. Αναλύωντας την συχνότητα εμφάνισης των λέξεων μπορούμε να υπολογίσουμε το “βάρος” τους.

Το βάρος μίας λέξης μετράει την ικανότητα της λέξης να διακρίνει το κείμενο στο οποίο ανήκει από τα υπόλοιπα κείμενα μίας συλλογής. Μπορεί επίσης να είναι ένα μέτρο του πόσο η συγκεκριμένη λέξη σχετίζεται με τα ενδιάφεροντα του χρήστη.

Το βάρος των λέξεων μας επιτρέπει να είμαστε πιο ακριβείς κατά την ανάκτηση της πληροφορίας. Επίσης, αν περιοριστούμε στις πιο σημαντικές λέξεις μειώνουμε σημαντικά τις διαστάσεις του χώρου στον οποίο πραγματοποιούμε τους υπολογισμούς μας.

A Bag of Words

the price of Words

Το AdWords είναι η pay-per-click διαφημιστική υπηρεσία της Google και η βασική πηγή εσόδων της εταιρείας. Οι διαφημίσεις της Google παρουσιάζονται υπό την μορφή σύντομου κειμένου με συνοπτικό τίτλο και δύο γραμμές περιεχομένου. Διαφέρουν εμφανισιακά ελάχιστα από τα αποτελέσματα μίας αναζήτησης και σχετίζονται με το περιεχόμενό τους. Οι διαφημιζόμενοι πληρώνουν για τις λέξεις κλειδιά με τις οποίες οι χρήστες αναζητούν πληροφορίες. Το κόστος μίας λέξης αυξάνεται με την συχνότητα χρήσης της.

Το AdSence είναι διαφημιστική υπηρεσία της Google που επιτρέπει σε ιδιοκτήτες ιστοχώρων να προσθέσουν σε αυτούς διαφημίσεις. Η επιλογή των διαφημίσεων γίνεται με βάση το περιεχόμενο της σελίδας. Ποσοστό από τα έσοδα των διαφημίσεων πηγαίνει στον ιδιοκτήτη του ιστοχώρου.

A Bag of Words