ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ
-
Εισαγωγή στην εξόρυξη δεδομένων

Καθ. Μ. Βαζιργιάννης,
Ερευνητική Ομάδα Εξόρυξης Γνώσης
Dept of Informatics,
Athens Univ. of Economics & Business
Patision 76, 10434, Athens, Greece
www: http://www.db-net.aueb.gr/

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ – DATA MINING

- Εισαγωγη στην Εξορυξη Δεδομενων
- Κινητρα για data mining
- Περιοχες εφαρμογων της τεχνολογιας αυτης
- Περιοχες κρατικης δραστηριοτητας που θα ωφελουνταν απο την χρηση και εμπεδωση τετοιων τεχνολογιων
- Αναφορα σε σχετικα εργαλεια ανοικτου λογισμικου που χρησιμοποιουμε
- Παρουσιαση καποιων δραστηριοτητων του εργαστηριου μας στον χωρο αυτο (px. gov.gr web archive, biomedical antimicrobial resistance surveylance, web mining DIGITEO grant...)
- λιγα λογια για το διεθνες συνεδριο ECML/PKDD 2011 που οργανωνουμε στην Αθηνα Σεπτ. 2011.

Εισαγωγή στην Εξόρυξη Δεδομένων

- Τι είναι η Εξόρυξη Δεδομένων;
- Τα σύνολα των δεδομένων
  - Το “data matrix”
  - Άλλα σχήματα δεδομένων
- Σκοπός της Εξόρυξης δεδομένων
  - Πρόβλεψη και περιγραφή
- Αλγόριθμοι της Εξόρυξης Δεδομένων
  - Συναρτήσεις αποτελεσμάτων, μοντέλα, και μέθοδοι βελτιστοποιήσης

Τι είναι η Εξόρυξη Δεδομένων;

Τι είναι Εξόρυξη Δεδομένων;

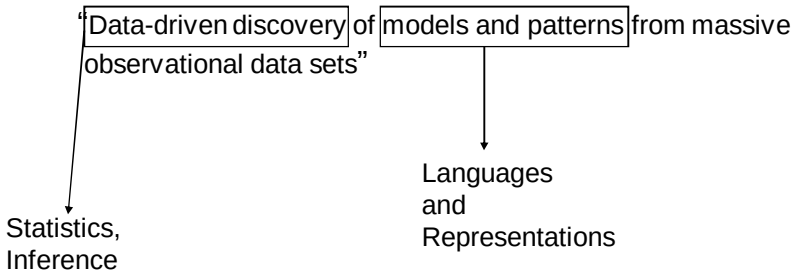
"Data-driven discovery of models and patterns from massive observational data sets"

Τι είναι Εξόρυξη Δεδομένων;

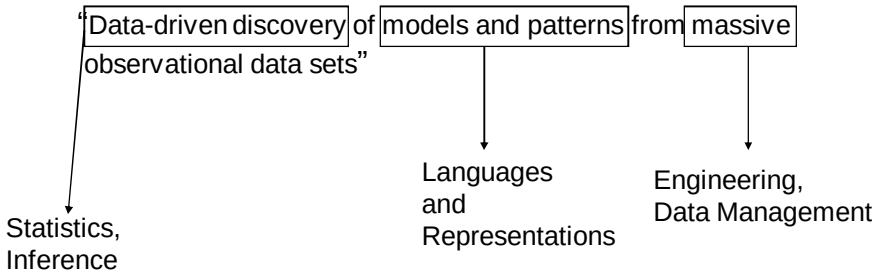
"Data-driven discovery of models and patterns from massive observational data sets"

Statistics,
Inference

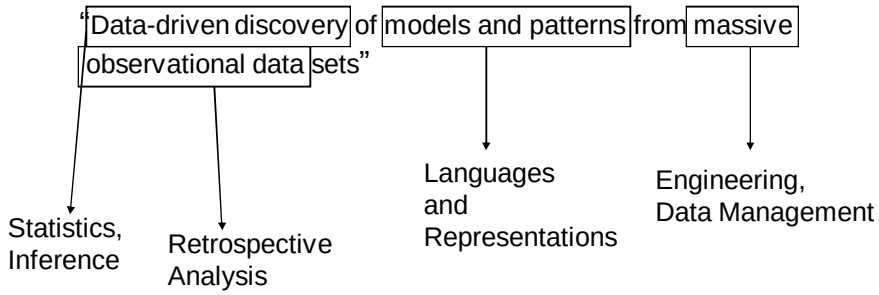
Τι είναι Εξόρυξη Δεδομένων;



Τι είναι Εξόρυξη Δεδομένων;



Τι είναι Εξόρυξη Δεδομένων;



Τεχνολογικοί παράγοντες

- Μεγαλύτερη, φθηνότερη μνήμη
  - Ο νόμος του Moore για τη μαγνητική πυκνότητα του δίσκου "διπλάσια ικανότητα κάθε 18 μήνες"
  - Κόστος Αποθήκευσης ανά byte που πέφτει γρήγορα
- Γρηγορότεροι, φθηνότεροι επεξεργαστές
  - the CRAY of 15 years ago is now on your desk
- Επιτυχία στην τεχνολογία συγγένειας των βάσεων
  - Ο Καθένας είναι ένας «ιδιοκτήτης» βάσεων δεδομένων
- Καινούριες ιδέες στη μηχανική μάθηση/στατιστικές μέθοδοι
  - Boosting, SVMs, δέντρα αποφάσεων, etc

Παραδείγματα ογκωδών συλλογών δεδομένων

- MEDLINE: ιατρική βάση βιβλιογραφίας
  - 12 εκατ. δημοσιευμένα άρθα
- Google
  - >20 δισ-εκατ. ιστοσελίδες
  - ~200 εκατ. επισκέπτες την ημέρα
- CALTRANS δεδομένα από αισθητήρες
  - Κάθε 30 δευτερόλεπτα, χιλιάδες από αισθητήρες, συλλέγουν 2Gbytes την ημέρα
- NASA MODIS δορυφόρος
  - Κάλυψη σε 250m ανάλυσης, σε 37 ζώνες, όλη η γη κάθε μέρα
- Retail – Wallmart:
  - ~100 εκατ. Δοσοληψίες την ημέρα.

Δύο Τύποι δεδομένων

- **Πειραματικά Δεδομένα**
  - Υπόθεση Η
  - Σχεδίαση ενός πειράματος για έλεγχο της υπόθεσης Η
  - Συλλογή δεδομένων, συμπεραίνουμε πόσο πιθανή είναι η υπόθεση Η να είναι αληθή
  - π.χ., κλινικές δοκιμές στην ιατρική
- **Δεδομένα από Παρατηρήσεις, Αναδρομικά δεδομένα**
  - Δεδομένα που δεν έχουν παραχθεί από πείραμα
    - πχ. human genome, atmospheric simulations.
  - Πειραματικές υποθέσεις δεν υφίστανται πλέον
  - Πως μπορούμε να χρησιμοποιήσουμε τα δεδομένα στην επιστήμη?
    - Τα δεδομένα πρέπει να υποστηρίζουν έλεγχος υπόθεσης (hypothesis testing)

Data-Driven Discovery

- Δεδομένα παρατήρησης
  - Φθηνά σε σχέση με τα πειραματικά δεδομένα
    - Παραδείγματα:
      - Εξερεύνηση των αρχείων δοσοληψιών από τον χώρο retail, αερογραμμές , κτλ
      - Διαδικτυακά αρχεία καταγραφής (Web logs) Amazon, Google, κτλ
      - Το γονίδιωμα ανθρώπινο/ποντικιού
      - ...
    - ⇒ Έχει νόημα να κάνεις χρήση των διαθέσιμων δεδομένων.
    - ⇒ χρήσιμες (?) πληροφορίες μπορεί να κρύβονται σε τεράστια αρχεία δεδομένων

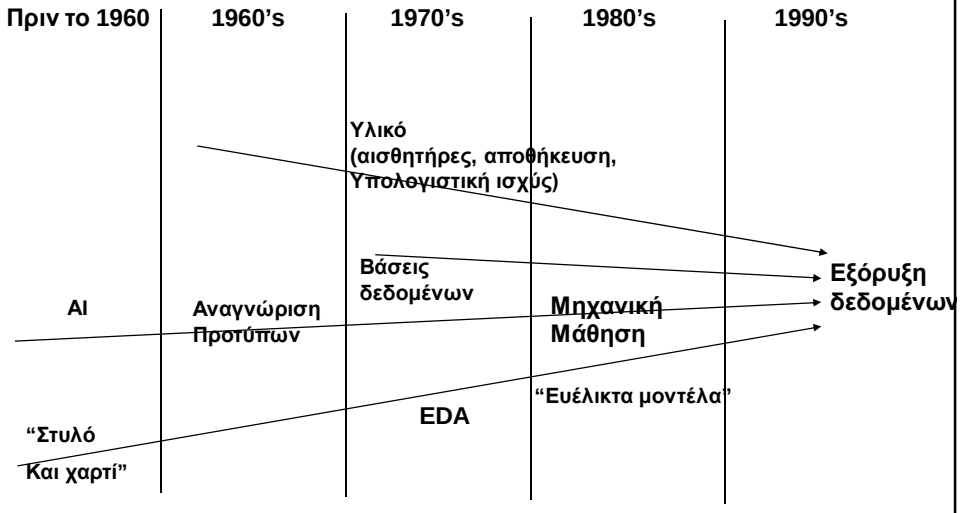
- Διαφορές της εξόρυξης δεδομένων με τις παραδοσιακές στατιστικές μεθόδους
  - Παραδοσιακές στατιστικές μέθοδοι: αρχικά η υπόθεση, έπειτα συλλογή δεδομένων, στη συνέχεια ανάλυση
  - Εξόρυξη Δεδομένων:
    - Λίγες η καθόλου αρχικές υποθέσεις,
    - Συνήθως τα δεδομένα υπάρχουν
    - Η ανάλυση είναι οδηγούμενη από τα δεδομένα όχι από την υπόθεση
  - Παρ' όλα αυτά, οι ιδέες στατιστικών μεθόδων είναι αρκετά χρήσιμες στην εξόρυξη δεδομένων, π.χ., Στατιστική αποτίμηση του κατά πόσο τα παραχθέντα μοντέλα είναι χρήσιμα

Αφήστε να μιλήσουν τα δεδομένα...

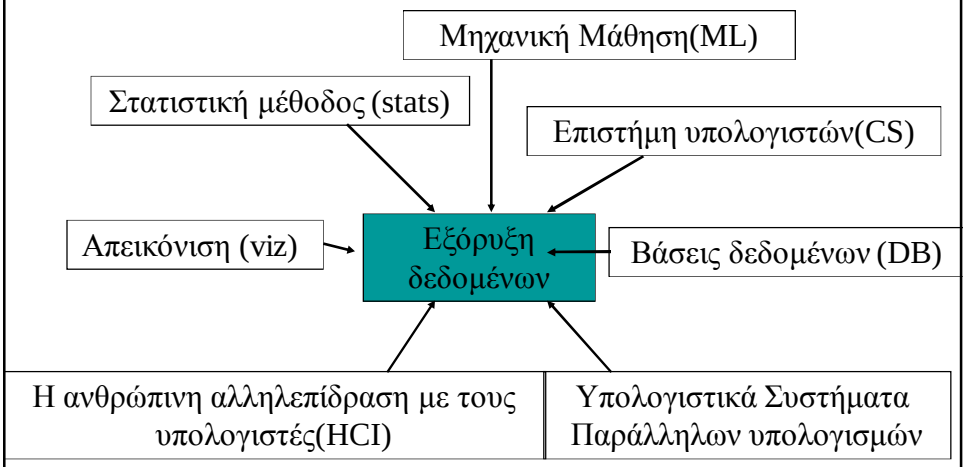
Αφήστε να μιλήσουν τα δεδομένα...

Τα δεδομένα μπορεί να έχουν πολλά να πουν ….. αλλά ίσως να είναι «θορυβώδη»!

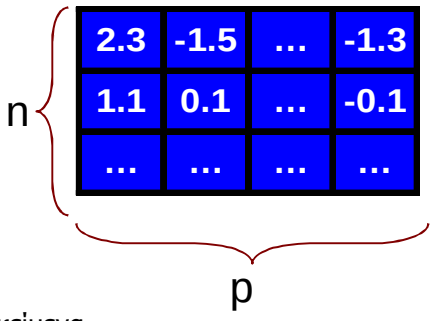
Η προέλευση της εξόρυξης δεδομένων



DM: Τομή επιστημονικών πεδίων

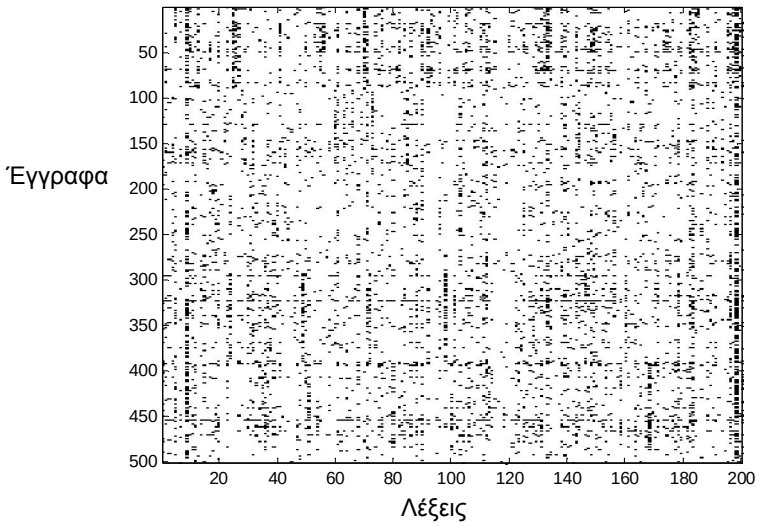


Επίπεδα αρχεία ή Διανύσματα δεδομένων

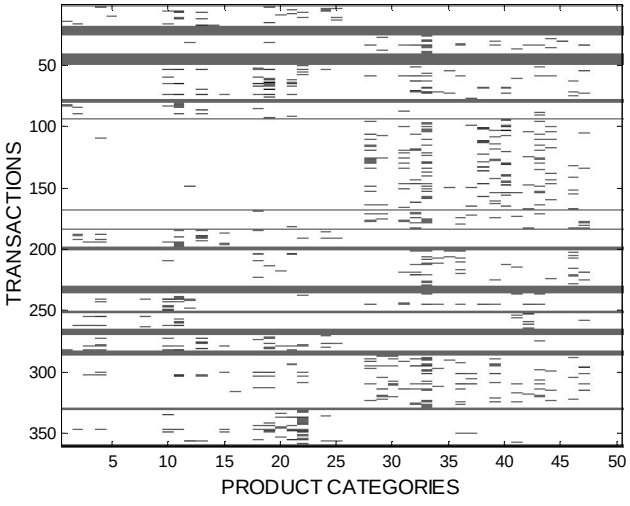


- Σειρές=αντικείμενα
- Στήλες= μετρήσεις πάνω στα αντικείμενα
  - Σειρά ως p-διαστάτο διάνυσμα, όπου p ο αριθμός των διαστάσεων.
    - έτσι έχουμε ενσωματωμένα τα αντικείμενα μας σε p-διαστάσεις διανυσματικού χώρου
- n και p μπορούν να είναι πολύ μεγάλα σε συγκεκριμένες εφαρμογές εξόρυξης δεδομένων

Sparse Matrix (κείμενο) δεδομένων



"Market Basket" Data

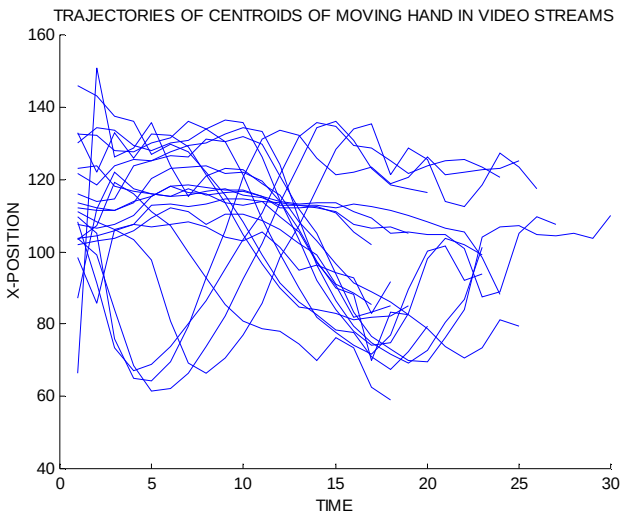


Δεδομένα Χρήσης από το Web

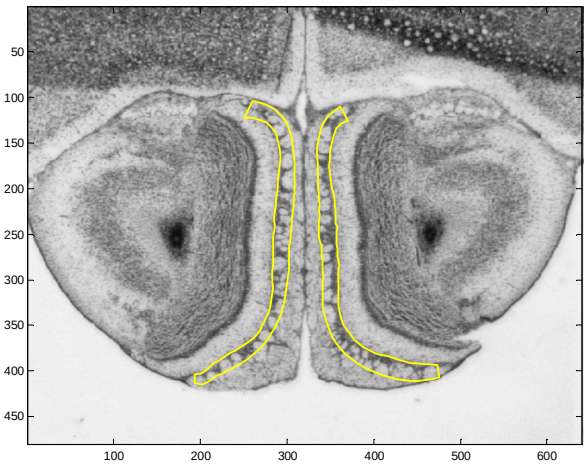
128.195.36.195, -, 3/22/00, 10:35:11, W3SVC, SRVR1, 128.200.39.181, 781, 363, 875, 200, 0, GET, /top.html, -,
128.195.36.195, -, 3/22/00, 10:35:16, W3SVC, SRVR1, 128.200.39.181, 5288, 524, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.195, -, 3/22/00, 10:35:17, W3SVC, SRVR1, 128.200.39.181, 30, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.195.36.101, -, 3/22/00, 16:18:50, W3SVC, SRVR1, 128.200.39.181, 60, 425, 72, 304, 0, GET, /top.html, -,
128.195.36.101, -, 3/22/00, 16:18:58, W3SVC, SRVR1, 128.200.39.181, 8322, 527, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.101, -, 3/22/00, 16:18:59, W3SVC, SRVR1, 128.200.39.181, 0, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:54:37, W3SVC, SRVR1, 128.200.39.181, 140, 199, 875, 200, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 17766, 365, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:07, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 1061, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:39, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:03, W3SVC, SRVR1, 128.200.39.181, 1081, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:56:04, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:33, W3SVC, SRVR1, 128.200.39.181, 0, 262, 72, 304, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:56:52, W3SVC, SRVR1, 128.200.39.181, 19598, 382, 414, 200, 0, POST, /spt/main.html, -,

Χρήστης 1	2	3	2	2	3	3	3	1	1	1	3	1	3	3	3	3
Χρήστης 2	3	3	3	1	1	1										
Χρήστης 3	7	7	7	7	7	7	7	7								
Χρήστης 4	1	5	1	1	1	5	1	5	1	1	1	1	1	1		
Χρήστης 5	5	1	1	5												
...	...															

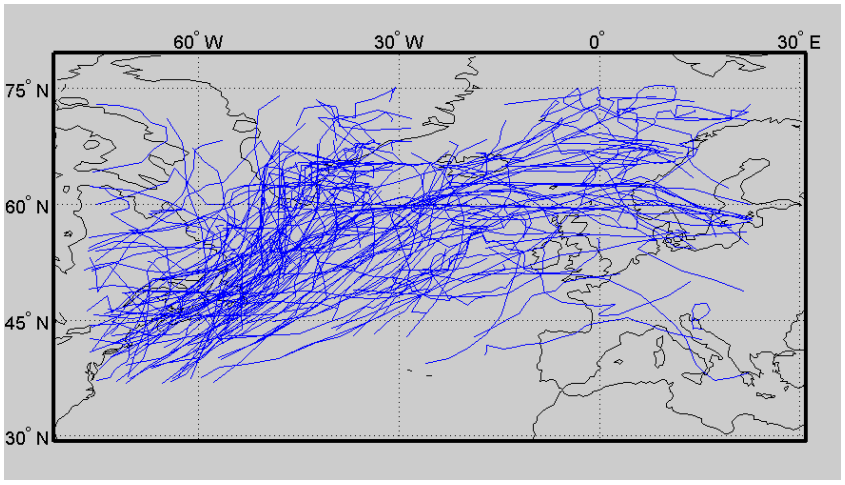
Χρονοσειρές δεδομένων

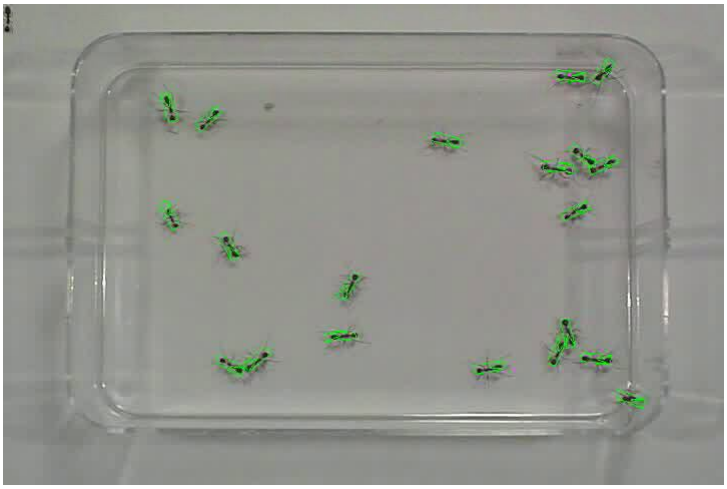


Εικόνες



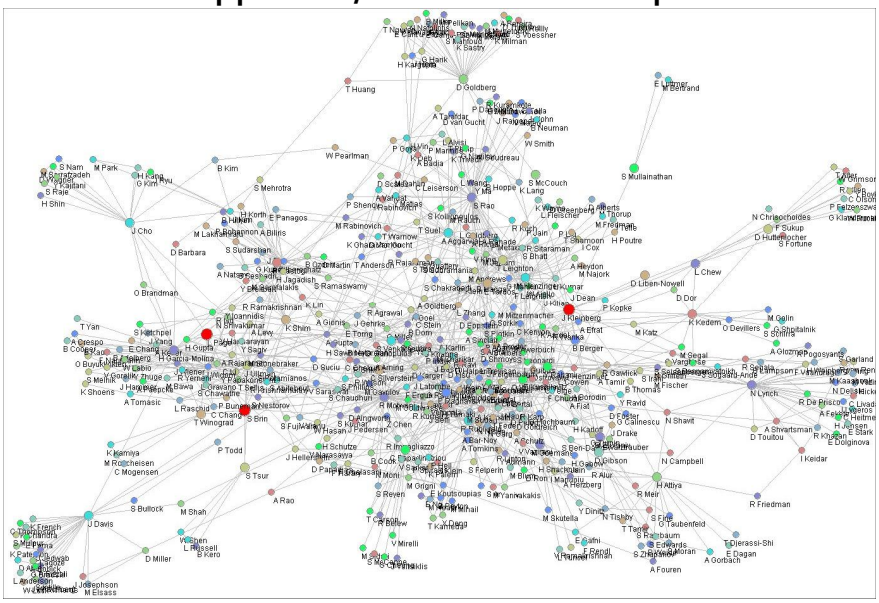
Χωροχρονικά δεδομένα



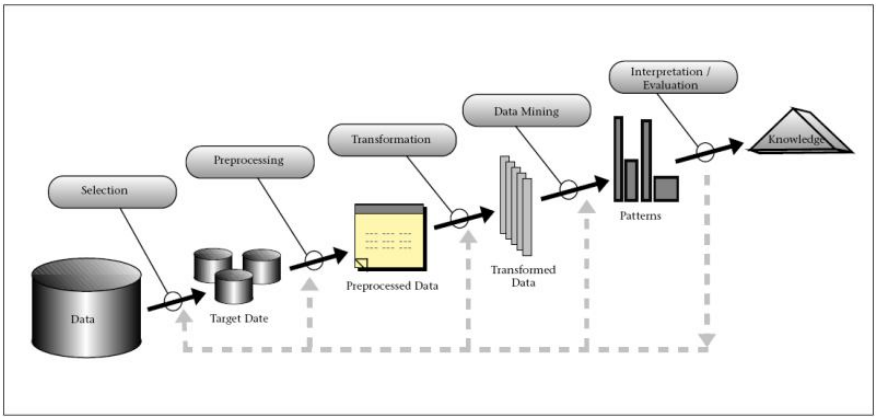


Tucker Balch and Frank Dellaert,
Τμήμα Επιστήμης Υπολογιστών, Georgia Tech

Συγγενικοί/Κοινωνικοί Δεσμοί



Διεργασίες εξόρυξης δεδομένων



Εργασίες της εξόρυξης δεδομένων

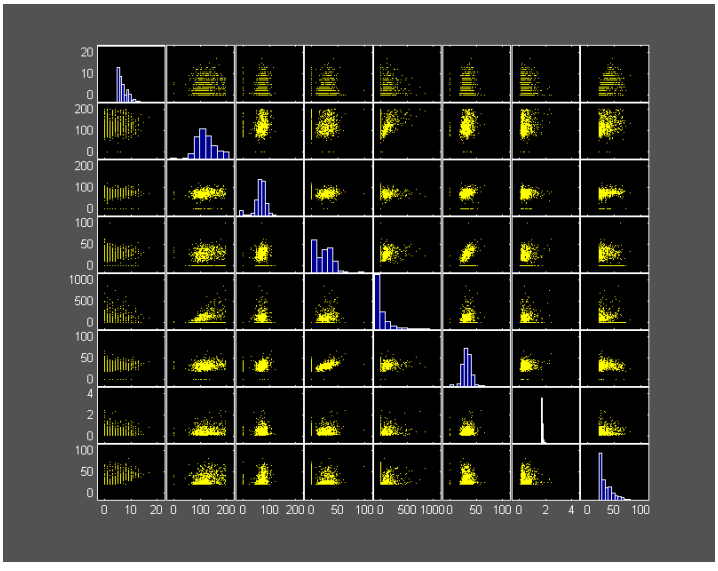
- Διερευνητική ανάλυση των δεδομένων (exploratory data analysis)
- Μοντέλο για περιγραφική αναπαράσταση (descriptive data modeling)
- Μοντέλο για προβλέψεις (predictive modeling)
- Ανακαλύπτοντας Μοτίβα και Κανόνες (patterns and rules)
- + άλλα….

Διερεύνηση των δεδομένων (Exploratory Data Analysis)

- Εύρεση μιας συνολικής άποψης των δεδομένων
  - Υπολογισμός συνολικών στατιστικών:
    - Πλήθος διαφορετικών τιμών, μέγιστο, ελάχιστο, μέσος, διασπορά, παραμόρφωση κατανομής κλπ.
- Μέθοδοι απεικόνισης χρησιμοποιούνται ευρέως
  - 1d ιστογράμματα
  - 2d scatter plots
  - Higher-dimensional methods
- Χρήσιμο για έλεγχο δεδομένων
  - E.g., βρίσκουμε ότι μία μεταβλητή είναι πάντα ένας ακέραιος αριθμός εκτιμημένος ή θετικός
  - Βρίσκουμε κάποιες μεταβλητές είναι ισχυρά πολωμένες (skewed)
- Απλές τεχνικές για διερευνητική ανάλυση μπορεί να είναι εξαιρετικά πολύτιμες
  - Πρέπει πάντα "να κοιτάς" τα δεδομένα σου πριν εφαρμόσεις οποιοδήποτε αλγόριθμο εξόρυξης δεδομένων.

Παραδείγματα της διερεύνησης δεδομένων

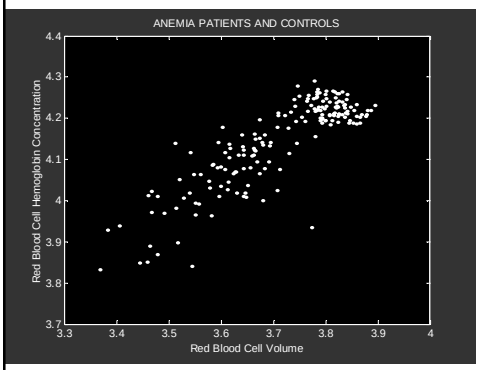
(Pima Indians data, scatter plot matrix)

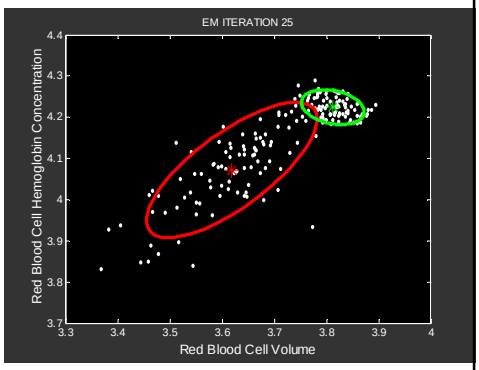


Μοντέλα για περιγραφική αναπαράσταση

- Στόχος είναι ένα "παραγωγικό" ή "περιγραφικό" μοντέλο,
  - Π.χ., ένα μοντέλο που μπορούσε να μιμηθεί τα δεδομένα αν χρειαζόταν
  - Αναπαριστά την διαδικασία που παράγει τα δεδομένα
- Παραδείγματα:
  - Εκτίμηση πυκνότητας:
    - Εκτίμηση κοινής κατανομής $P(x_1, ...x_p)$
  - Cluster analysis:
    - Βρίσκει τις φυσικές ομάδες δεδομένων
  - Εξάρτηση των μοντέλων μεταξύ των p μεταβλητών
    - Μαθαίνουμε ένα Bayesian δίκτυο για τα δεδομένα

Παράδειγμα περιγραφικού μοντέλου

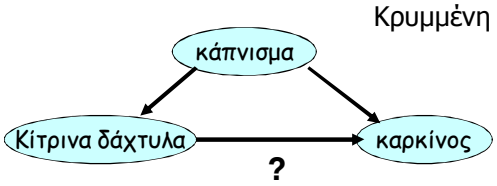




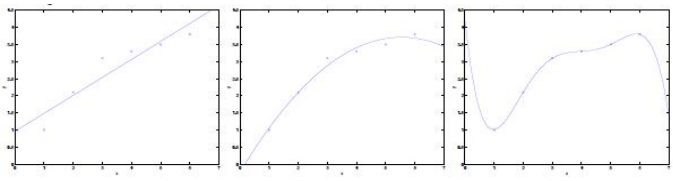
Αλλο μοντέλο περιγραφικής αναπαράστασης

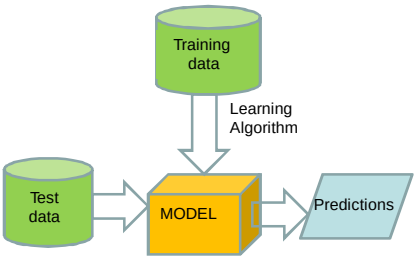
- Μαθαίνοντας κατευθυνόμενα γραφικά μοντέλα (aka Bayes Nets)
  - Στόχος: να μάθουμε τις κατευθυνόμενες σχέσεις μεταξύ p μεταβλητών
  - Τεχνικές: κατευθυνόμενες (causal) γράφοι
  - Πρόκληση: Διάκριση μεταξύ συσχέτισης και αιτιότητας
    - παράδειγμα: Προκαλούν τα κίτρινα δάχτυλα καρκίνο στους πνεύμονες;

Κρυμμένη αιτία: κάπνισμα



Μοντέλο για προβλέψεις





- Με βάση ένα σύνολο μεταβλητών X προβλέπει μία μεταβλητή Y
  - X μπορεί να είναι p-διαστάσεων διάνυσμα
- Ταξινόμηση: Y λεκτική τιμή (πχ. «απάτη», «όχι απάτη»
- Παλινδρόμηση: Y αριθμητική τιμή (πχ. Τιμή μιας μετοχής)
- Στην πραγματικότητα είναι η μάθηση της σχέσης μεταξύ Y και X
- Αλγόριθμοι: Decision Trees, Regression, Naïve Bayes, SVMs, ...
- Συχνά δίνεται έμφαση στην ακρίβεια των προβλέψεων, και λιγότερη σημασία στην κατανόηση των μοντέλων

Παραδείγματα των μοντέλων πρόβλεψης

- Υπόβαθρο
  - AT&T έχει περίπου 100 εκατ. πελάτες
  - Καταγράφονται ~200 εκατ. κλήσεις την ημέρα, 40 ιδιότητες το καθένα
  - 250 εκατ. μοναδικοί τηλεφωνικοί αριθμοί κλήσης
  - *Ποιοι αντιστοιχούν σε «εταιρίες» και ποιοι σε «κατοικίες»* ;
- Λύση (Pregibon and Cortes, AT&T,1997)
  - Μοντέλα πρόβλεψης, εκπαιδεύονται σε λίγες ιδιότητες με βάση γνωστά δεδομένα εταιρικών πελατών και αποδίδουν την πιθανότητα: p(εταιρία|δεδομένα)
  - Σημαντική υποδομή: τα δεδομένα αντιγράφονται κάθε βράδυ, ενημέρωση μοντέλου πρόβλεψης (20 επεξεργαστές, 6Gb RAM, terabyte disk farm)
  - Τρέχει καθημερινά

Ανεύρεση προτύπων/μοτίβων

- Στόχος είναι να ανακαλύψουμε “τοπικά” μοτίβα/συσχετίσεις μέσα στα δεδομένα και όχι να χαρακτηρίσουμε τα δεδομένα συνολικά
- Με βάση το «καλάθι αγορών» μπορέσαμε ν’ ανακαλύψουμε
  - οι πελάτες που αγοράζουν κρασί και ψωμί αγοράζουν τυρί με πιθανότητα 0.9
  - “Κανόνες συσχέτισης”
- Έστω αστρονομικά αντικείμενα με πολλές μεταβλητές
  - Πιθανόν να βρούμε μία μικρή ομάδα από προηγούμενα ανεξερεύνητα αντικείμενα έχουν μεγάλη ομοιότητα μεταξύ τους και είναι πολύ ανόμοια με όλα τα άλλα αντικείμενα.

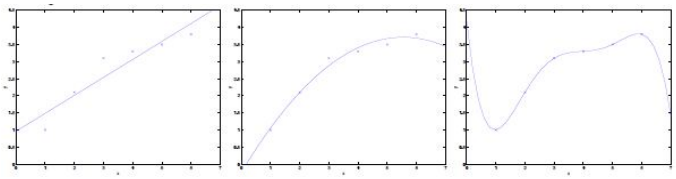
Παράδειγμα ανεύρεσης μοτίβων

ADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADAADABDBBDABABBCDD
DCDDABDCBBDBDBCBBABBBCBBABCBBACBBDBAACCADDADBDBBCBBCCBB
BDCABDDBBADDBBBBCCACDABBABDDCDDBBABDBDDBDDBCACDBBCCBBAC
DCADCBACCADCCCACCDDADCBCADADBAACCDDDCBDBDCCCCACACACCDAB
DDBCADADBCBDDADABCCABDAACABCABACBDDDCBADCBDADDDDCDDCADC
CBBADABBAAADAAABCCBCABDBAADCBCDACBCABABCCBACBDABDDDADAA
BADCDCCDBBCDBDADDCCBBCDBAADADBCAAAADBDCADBDBBBCDCCBCCCD
CCADAADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADACCCDABAABBC
BDBDBADBBBBCDADABABBDACDCDDDBBCDBBCBBCCDABCADDADBACBBBC
CDBAAADDDBDDCABACBCADCDCBAAADCADDADAABBACCBB

Παράδειγμα ανεύρεσης μοτίβων (βιολογικά δεδομένα)

ADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADAADABDBBDABABBCDD
DCDDABDCBBDBDBCBBABBBCBBABCBBACBBDBAACCADDADBDBBCBBCCBB
BDCABDDBBADDBBBBCCACDABBABDDCDDBBABDBDDBDDBCACDBBCCBBAC
DCADCBACCADCCCACCDDADCBCADADBAACCDDDCBDBDCCCCACACACCDAB
DDBCADADBCBDDADABCCABDAACABCABACBDDDCBADCBDADDDDCDDCADC
CBBADABBAAADAAABCCBCABDBAADCBCDACBCABABCCBACBDABDDDADAA
BADCDCCDBBCDBDADDCCBBCDBAADADBCAAAADBDCADBDBBBCDCCBCCCD
CCADAADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADACCCDABAABBC
BDBDBADBBBBCDADABABBDACDCDDDBBCDBBCBBCCDABCADDADBACBBBC
CDBAAADDDBDDCABACBCADCDCBAAADCADDADAABBACCBB

Δομή: Μοντέλα και Μοτίβα



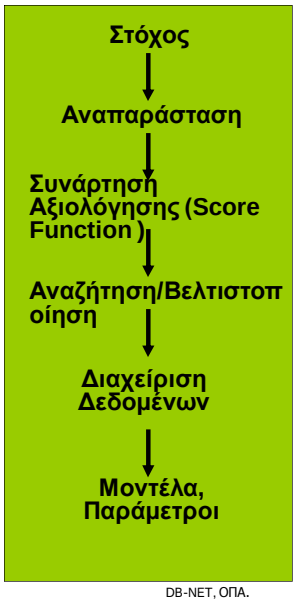
- Μοντέλα = αφηρημένη αναπαράσταση μιας διαδικασίας.
  π.χ., Η απλούστερη γραμμική μορφή δομής

  $$Y = a X + b$$

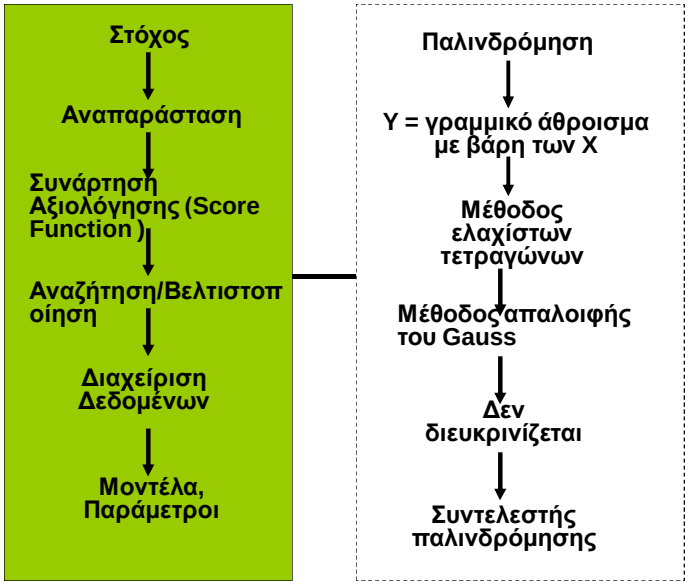
  - a και b είναι οι παράμετροι που υπολογίζονται από τα δεδομένα
  - Y = aX + b είναι η δομή του μοντέλου
  - Y = 0.9X + 0.3 είναι ένα συγκεκριμένο μοντέλο
  - "Όλα τα μοντέλα είναι λάθος, κάποια είναι χρήσιμα" (G.E. Box)

Στοιχεία Αλγορίθμων Data Mining

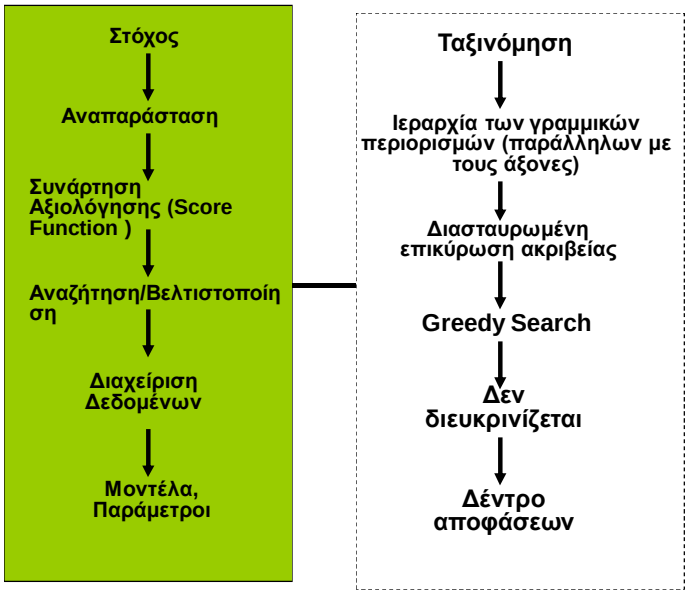
- Αναπαράσταση:
  - Αποφασίζοντας τη φύση και τη δομή της αναπαράστασης που χρησιμοποιείται
- Συνάρτηση Αξιολόγησης (Score function)
  - ποσοτικοποίηση και σύγκριση για το πόσο καλά διαφορετικές αναπαραστάσεις ταιριάζουν στα δεδομένα.
- Μέθοδος βελτιστοποίησης για αναζήτηση βέλτιστης λύσης
  - Επιλογή αλγοριθμικής διαδικασίας για να βελτιστοποιήσουμε το αποτέλεσμα λειτουργίας
- Διαχείριση αποτελεσμάτων
  - Ποιες αρχές της διαχείρισης των δεδομένων απαιτούνται για να εφαρμόσουμε τους αλγόριθμους αποτελεσματικά.



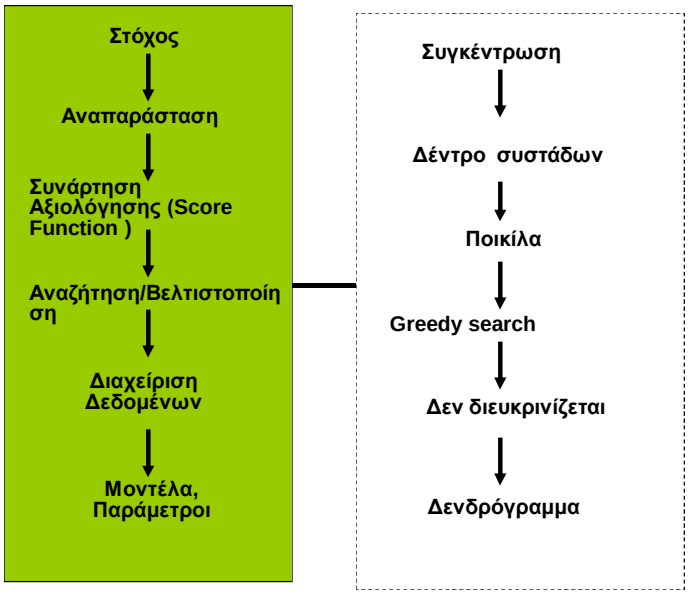
Ένα παράδειγμα: Γραμμική παλινδρόμηση με πολλές μεταβλητές



Ένα παράδειγμα: Δέντρο αποφάσεων (C4.5 or CART)



Ένα παράδειγμα: Ιεραρχική Συσταδοποίηση



Τομείς δημόσιου χώρου για εφαρμογή αλγορίθμων εξόρυξης δεδομένων

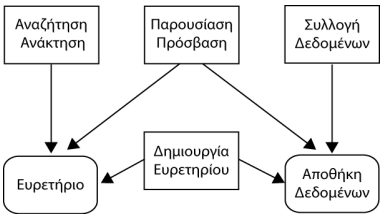
- TAXIS – *ELENXIS: για ανίχνευση patterns παραβατικής συμπεριφοράς με χρήση αλγορίθμων επιβλεπομένης μάθησης*
- *Public Web content (πχ ΔΙΑΥΓΕΙΑ, opengov): ανάγκη για αρχειοθέτηση και αναζήτηση στο ιστορικό περιεχόμενο.*
- *Δεδομένα προμηθειών: data warehousing και αναζήτηση προϊόντων /προσφορών, ομοιότητες*
- *Νομικά κείμενα: text mining & matching σε κείμενα νόμων, νομολογίες, αποφάσεις δικαστηρίων, εφημερίδας κυβέρνησης*

Δραστηριότητες DB-NET σε data/web mining

- Web Archiving & Indexing (gov.gr web archive)
- Web mining - DIGITEO grant -France
- Automated keyword extraction for web advertizing campaigns

Web Archiving & Indexing (gov.gr web archive)

Crawling & Indexing ~2300 governmental web sites



Technologies Used

Crawler

Heritrix 3 – open source provided by Internet Archive (IA) under LGPL.
Based on the *International Internet Preservation Consortium (IIPC) specs – 2003*
- Data crawled are stored in WARC (Web ARChive) format.

WayBack engine

- Για την παρουσίαση και πρόσβαση των δεδομένων
- *open source* υλοποίηση σε Java της *Internet Archive WayBack Machine*.
- είναι υπεύθυνο για την δημιουργία του ευρετηρίου
- παρέχει δυνατότητες αναζήτησης και ανάκτησης.
-Προς το παρόν η αναζήτηση γίνεται μόνο μέσω των URLs.
-Στη συνέχεια θα προστεθεί δυνατότητα αναζήτησης κειμένου (full-text search).
- Η μηχανή είναι προσβάσιμη μετά από συνεννόηση…
-Ένα πλήρως ανεπτυγμένο πρωτότυπο για τις ιστοσελίδες του Οικονομικού Πανεπιστήμιου είναι στο: http://archive.aueb.gr/

Recent publication

- Vassilis Plachouras, Chrysostomos Kapetis, Michalis Vazirgiannis, "Archiving the Web sites of Athens University of Economics and Business", στο 19ο Πανελλήνιο Συνέδριο Ακαδημαϊκών Βιβλιοθηκών

DIGITEO Chair Grant – funded by the DIGITEO alliance in France

Objectives

- Predictions in the Web
- Detection & Evaluation of Communities in Social networks & citation graphs
- Real time recommendations

Recent Publications:

-C. Giatsidis, D. Thilikos, M. Vazirgiannis, "Evaluating cooperation in communities with the k-core structure", in the proceedings of the 2011 IEEE - International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Taiwan.

- M. Rallis, M. Vazirgiannis, "Rank Prediction in graphs with Locally Weighted Polynomial Regression and EM of Polynomial Mixture Models", in the proceedings of the 2011 IEEE - International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Taiwan.

- Fully fledged Demo available at:
http://www.lix.polytechnique.fr/~giatsidis/

Automated keyword extraction for web advertizing campaigns

Motivation

-Online advertising is a very profitable industry

- development of ad campaigns is a laborious task involving significant human resources and expertise.

Objectives

-a system for multiword keyword recommendations in a semiautomatic manner.

What it does…

-Given a landing page, the system extracts relevant terms consisted of two or three words to match a potential search query.

- it proposes the most relevant keywords and other suggested terms that do not exist in the landing page text using search result snippets.

Performance

-blind testing experiments on real world data indicate that our approach performs equally to prominent existing industrial solutions in most of the cases.

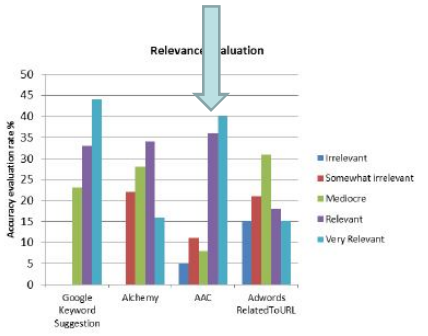


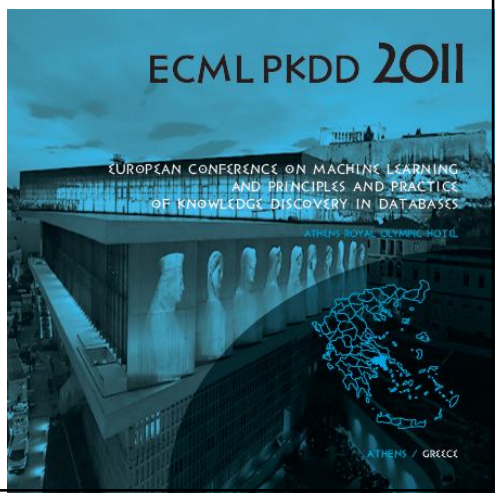
Figure 1. Answers of human evaluators reviewing the output of each system (relevance of keywords)

Recent Publication

- M. Thomaidou, M. Vazirgiannis, "Multiword Keyword Recommendation System for Online Advertising", in the proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Taiwan.

ΕΥΧΑΡΙΣΤΩ ΓΙΑ ΤΗΝ ΠΡΟΣΟΧΗ ΣΑΣ

http://www.db-net.aueb.gr/michalis/



ΠΡΟΣΚΛΗΣΗ ΝΑ ΣΥΜΜΕΤΑΣΧΕΤΕ στο European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) will take place in Athens, Greece from September 5th to 9th, 2011 http://www.ecmlpkdd2011.org/