Εισαγωγή στην Εξόρυξη Γνώσης

Ιωάννης Μαδεμλής

Ανακάλυψη Γνώσης

- **Ανακάλυψη γνώσης** (knowledge discovery) ονομάζεται ο κλάδος της τεχνητής νοημοσύνης που ασχολείται με την αυτοματοποιημένη αναζήτηση μοτίβων, τα οποία πιθανώς ανταποκρίνονται σε κάποια κριτήρια, μέσα σε μεγάλους όγκους δεδομένων.

Ανακάλυψη Γνώσης

- Η ανακάλυψη γνώσης αναλύει ένα σύνολο ακατέργαστων στοιχείων και εξάγει από αυτό νέα, μη τετριμμένη πληροφορία.

Ανακάλυψη Γνώσης

- Πηγές των ακατέργαστων δεδομένων μπορεί να είναι βάσεις δεδομένων, αρχεία κειμένου, ο Παγκόσμιος Ιστός, συστήματα λογισμικού, κλπ.
- Περιλαμβάνει μία αλληλουχία βημάτων: *επιλογή δεδομένων*, *προεπεξεργασία δεδομένων*, *εξόρυξη γνώσης*, *μετεπεξεργασία γνώσης* (αξιολόγηση και ερμηνεία των αποτελεσμάτων).
- Άρα, η **εξόρυξη δεδομένων/εξόρυξη γνώσης** (data mining) είναι ένα στάδιο στην αλληλουχία βημάτων της ανακάλυψης γνώσης σε σύνολα δεδομένων.

Ανακάλυψη Γνώσης

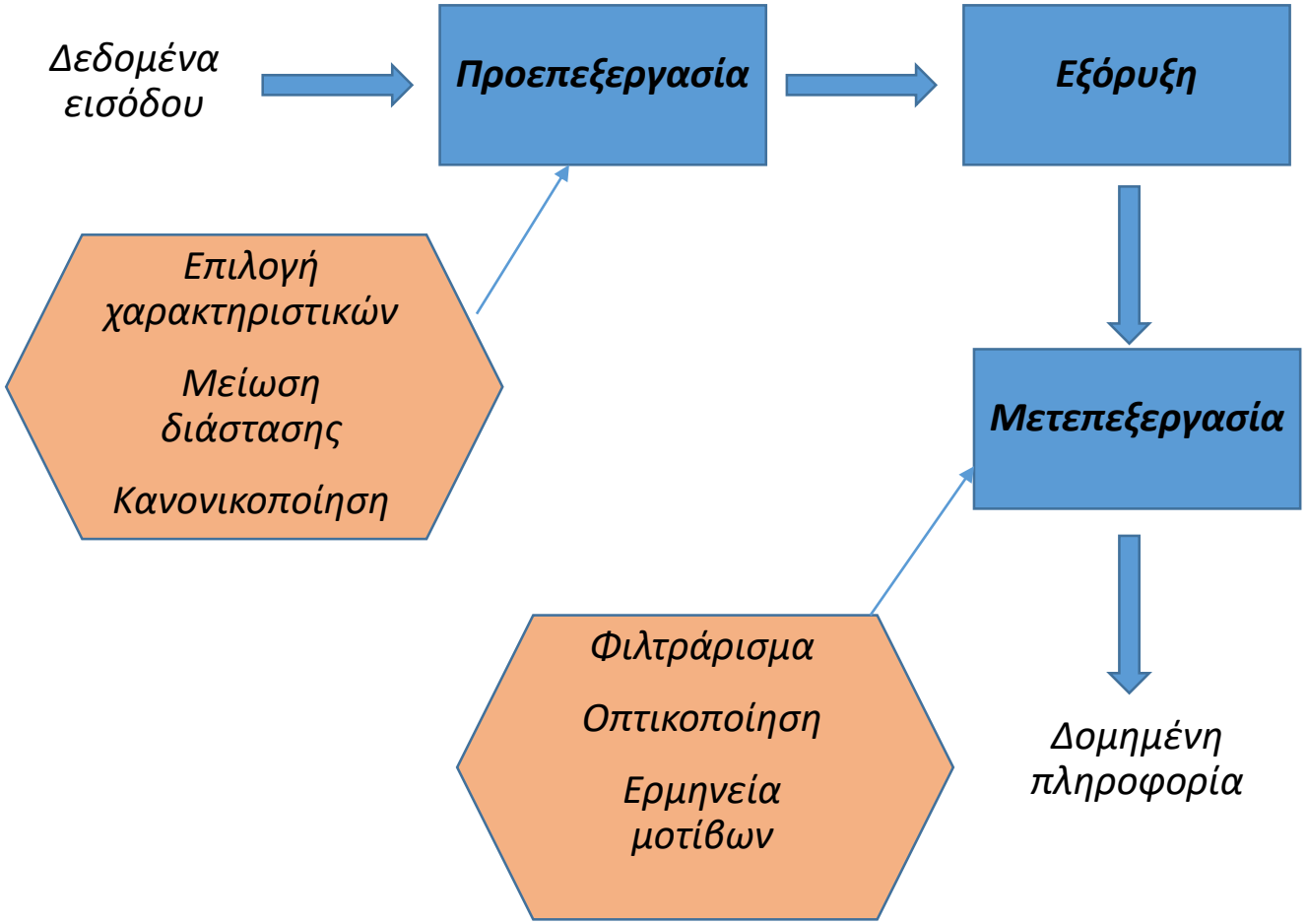
Εξόρυξη Γνώσης

- Το ορθότερο όνομα θα ήταν **εξόρυξη γνώσης από δεδομένα**.
- Η εξόρυξη γνώσης/δεδομένων βασίζεται σε μεθόδους από *τεχνητής νοημοσύνης* και *εφαρμοσμένης στατιστικής*.
- Η *εξόρυξη λογισμικού* είναι μία ειδική περίπτωση εφαρμοζόμενη σε σύνολα μεταδεδομένων που αφορούν κάποιο σύστημα λογισμικού (π.χ., αφαιρετικές μοντελοποιήσεις βάσεων δεδομένων).

Εξόρυξη Γνώσης

Εξόρυξη Γνώσης

Εξόρυξη Γνώσης



Εξόρυξη Γνώσης

- Η εξόρυξη γνώσης αξιοποιεί μεθόδους/αλγορίθμους της *αναγνώρισης προτύπων* ή *μηχανικής μάθησης.*
- Έχει όμως εφαρμοσμένο χαρακτήρα και περισσότερο ανεπίβλεπτο (unsupervised) προσανατολισμό.
- Δεν πρέπει να συγχέεται με την *ανάκτηση πληροφορίας* (information retrieval), η οποία εστιάζει στην εύρεση πληροφοριών σχετικών με κάποιο ερώτημα μέσα σε έναν μεγάλο όγκο αποθηκευμένων πληροφοριών.

Εξόρυξη Γνώσης

Προεπεξεργασία δεδομένων

- Η **προεπεξεργασία των δεδομένων** συνήθως μελετάται από κοινού με τους αλγορίθμους εξόρυξης.

- Τα βήματα αυτά είναι συνήθως απαραίτητα έτσι ώστε να δοθούν κατάλληλα δεδομένα ως είσοδοι στον αλγόριθμο.

- Περιλαμβάνουν επιλογή χαρακτηριστικών, συγχώνευση δεδομένων από διαφορετικές πηγές, καθαρισμό των δεδομένων (π.χ., αποθορυβοποίηση), αναπαράσταση των δεδομένων, κλπ.

Εξόρυξη Γνώσης

Μετεπεξεργασία δεδομένων

- Μετά την εξόρυξη, η αποκτηθείσα γνώση μπορεί να υποστεί **μετεπεξεργασία**, ώστε μόνο τα έγκυρα και χρήσιμα συμπεράσματα να αξιοποιηθούν.

- Πρόκειται για ένα σύνολο μεθόδων με στόχο:
  - Τη διερεύνηση των δεδομένων και των αποτελεσμάτων της εξόρυξης, και
  - Τον υπολογισμό στατιστικών μέτρων ή ελέγχου υποθέσεων για την εξάλειψη ενδεχομένως εσφαλμένων συμπερασμάτων.

Εξόρυξη Γνώσης

Εξόρυξη Γνώσης

Στόχοι Εξόρυξης Γνώσης

- Οι παραδοσιακές μέθοδοι ανάλυσης δεδομένων (π.χ., στατιστικές) ήταν πλέον ανεπαρκείς κατά τη δεκαετία του 1990.
  - Ως αποτέλεσμα, σχηματίστηκε η εξόρυξη γνώσης ώστε να αντιμετωπίσει τα εν λόγω ζητήματα.
- Γιατί; Προβλήματα:
  - α) **Κλιμακωσιμότητας**: τα σημερινά σύνολο δεδομένων μπορεί να είναι τεράστια όσον αφορά το πλήθος και το μέγεθος των δεδομένων που περιέχουν.
  - β) **Διαστατικότητας**: αν κάθε δεδομένο αναπαρίσταται ως ένα K-διάστατο διάνυσμα και το K είναι πολύ μεγάλο, οι παραδοσιακοί στατιστικοί αλγόριθμοι ενδεχομένως να μην αποδίδουν καλά, ή να παρουσιάζουν ζητήματα υπολογιστικής πολυπλοκότητας.

Εξόρυξη Γνώσης

Στόχοι Εξόρυξης Γνώσης

- Γιατί; Προβλήματα:
  - γ) **Ετερογένειας**: τα σημερινά σύνολα δεδομένων μπορεί να συμπεριλαμβάνουν ποικιλόμορφους τύπους δεδομένων στην ίδια βάση (π.χ., εικόνες, κείμενο και γράφους).
  - δ) **Κατανομής**: το σύνολο δεδομένων είναι διασπαρμένο σε πολλαπλούς, φυσικά απομακρυσμένους υπολογιστές (ζητήματα ελαχιστοποίησης επικοινωνιακού φόρτου, ασφάλειας, συγκέντρωσης αποτελεσμάτων, κλπ.).
  - ε) **Μη παραδοσιακή ανάλυση**: τα σημερινά σύνολα δεδομένων συνήθως δεν προκύπτουν από μία προσεκτικά σχεδιασμένη διαδικασία συλλογής, στηριγμένης σε μία εκ των προτέρων διατυπωμένη στατιστική υπόθεση.
    - Συχνά πρόκειται για *καιροσκοπικά* στατιστικά δείγματα αντί για *τυχαία* δείγματα: δεν έγινε πραγματικά τυχαία δειγματοληψία, αλλά συλλέξαμε ό,τι μπορέσαμε να βρούμε.

Εξόρυξη Γνώσης

Προέλευση Εξόρυξης Γνώσης

- Για την αντιμετώπιση των ζητημάτων αυτών, σχηματίστηκε το πεδίο της εξόρυξης γνώσης κατά τη δεκαετία του 1990.

- Συνιστά σημείο επαφής μεταξύ της **συμβολικής τεχνητής νοημοσύνης**, της **αναγνώρισης προτύπων/μηχανικής μάθησης** και των **βάσεων δεδομένων**.

- Το πεδίο των κατανεμημένων συστημάτων παίζει επίσης υποστηρικτικό ρόλο όταν τα σύνολα δεδομένων είναι τεράστια.

Προέλευση Εξόρυξης Γνώσης

Εξόρυξη Γνώσης

- **Σύνοψη**: Ανεπάρκεια των παραδοσιακών στατιστικών μεθοδολογιών.
  - Δεν μπορούσαν να επεξεργαστούν σωστά τα γιγάντια και περίπλοκα σύνολα στοιχείων που ήταν πλέον διαθέσιμα χάρη στη απανταχού παρουσία δικτυωμένων υπολογιστών.
  - Οι παραδοσιακές μέθοδοι παρουσίαζαν προβλήματα *κλιμακωσιμότητας*, *αδυναμίας επεξεργασίας διανυσματικών συνόλων δεδομένων με πολύ υψηλή διάσταση*, *αδυναμίας επεξεργασίας συνόλων δεδομένων που αναμείγνυαν γνωρίσματα διαφορετικού τύπου*, κλπ.
- **Σημείωση**: Στην επιτυχή εξόρυξη νέας γνώσης, σημαντικοί είναι τόσο οι χρησιμοποιούμενοι αλγόριθμοι, όσο και τα ίδια τα δεδομένα.
  - Πρέπει να εμπεριέχουν μη τετριμμένη γνώση και να είναι αντιπροσωπευτικά του προς επίλυση προβλήματος.

Δεδομένα

Αναπαράσταση δεδομένων

- Συνήθως τα σύνολα δεδομένων είναι σύνολα διανυσμάτων καθορισμένης διάστασης.
  - Κάθε *στοιχείο/διάνυσμα/πρότυπο/δεδομένο* αποτελείται από μία πλειάδα **γνωρισμάτων** (attributes) και κάθε γνώρισμα μπορεί να λάβει μία από πολλαπλές τιμές: διακριτές, αν είναι **διακριτή** μεταβλητή, ή μη μετρήσιμα άπειρες, αν είναι **συνεχής** μεταβλητή.
  - Εναλλακτικά, κάθε στοιχείο / πρότυπο θα μπορούσε να είναι ένας γράφος, ή απλώς ένα διάνυσμα ανεξάρτητο από τα υπόλοιπα (π.χ., διαφορετικής διάστασης).
- Κάθε τέτοιο πρότυπο καλείται **διάνυσμα χαρακτηριστικών** (feature vector).
  - Είναι ευθύνη δική μας, ή ενός αλγορίθμου προεπεξεργασίας, να μετατρέψουμε τα αρχικά ακατέργαστα δεδομένα σε διανύσματα χαρακτηριστικών.

Προβλήματα

Προβλήματα εξόρυξης

- Οι εργασίες με τις οποίες καταπιάνεται η εξόρυξη δεδομένων είναι δύο ειδών:
  - **Προβλεπτικές**, όπου σκοπός είναι η πρόβλεψη της τιμής ενός συγκεκριμένου γνωρίσματος-στόχου κάποιου νέου στοιχείου με βάση τις τιμές των υπολοίπων γνωρισμάτων σε παλαιότερα γνωστά στοιχεία και στο ίδιο, και
  - **Περιγραφικές**, όπου στόχος είναι η εύρεση μοτίβων που συνοψίζουν τις σχέσεις μεταξύ των δοθέντων δεδομένων.
- **Σημείωση**: Η προβλεπτική μοντελοποίηση συμπίπτει με ό,τι ονομάζουμε στη μηχανική μάθηση «*επιβλεπόμενη μάθηση*».
  - Υλοποιείται με την εκπαίδευση ενός μοντέλου έτσι ώστε να ελαχιστοποιείται το σφάλμα μεταξύ προβλεπόμενων και πραγματικών τιμών στόχου, σε ένα σύνολο *γνωστών* δεδομένων εκπαίδευσης.

Προβλήματα

Προβλήματα εξόρυξης

- Τα κύρια προβλήματα της εξόρυξης γνώσης είναι τα παρακάτω:
  - Α) **Προβλεπτική μοντελοποίηση**: κατασκευή ενός μοντέλου για το γνώρισμα-στόχο, ως συνάρτηση των υπολοίπων γνωρισμάτων.
    - Υπάρχουν δύο είδη προβλεπτικής μοντελοποίησης: η **ταξινόμηση** (classification), όπου το γνώρισμα-στόχος είναι διακριτή μεταβλητή, και η **παλινδρόμηση** (regression), όπου το γνώρισμα-στόχος είναι συνεχής/πραγματική μεταβλητή.
    - Ο σκοπός είναι και στις δύο περιπτώσεις η μάθηση ενός μοντέλου που ελαχιστοποιεί το σφάλμα μεταξύ των προβλεπόμενων και των πραγματικών τιμών του γνωρίσματος-στόχου στα *νέα* πρότυπα.
    - Παράδειγμα ταξινόμησης είναι το πρόβλημα της αναγνώρισης του αν ένας χρήστης ηλεκτρονικού καταστήματος θα προβεί ή όχι στην αγορά ενός συγκεκριμένου προϊόντος (δύο δυνατές κλάσεις ως στόχος: αρνητική/θετική).
    - Παράδειγμα παλινδρόμησης είναι η πρόβλεψη των μελλοντικών τιμών μίας μετοχής.

Προβλήματα εξόρυξης

- Τα κύρια προβλήματα της εξόρυξης γνώσης είναι τα παρακάτω:
  - Β) **Ομαδοποίηση** (clustering)**:** αναγνώριση επιμέρους *ομάδων* (clusters) από παρεμφερή στοιχεία μες στο ολικό σύνολο δεδομένων, έτσι ώστε τα στοιχεία που ανήκουν στην ίδια ομάδα να είναι περισσότερο όμοια μεταξύ τους απ' ότι με πρότυπα άλλων ομάδων.
    - Προαπαιτούμενο είναι να έχει οριστεί επακριβώς η έννοια της *ομοιότητας* (ή, αντιστρόφως, της *απόστασης*) μεταξύ δύο στοιχείων/προτύπων.
    - Η έννοια αυτή εξαρτάται από το εκάστοτε πρόβλημα και τη φύση του συνόλου δεδομένων.

Προβλήματα

Προβλήματα εξόρυξης

- Τα κύρια προβλήματα της εξόρυξης γνώσης είναι τα παρακάτω:
  - Γ) **Ανάλυση συσχετίσεων** (association analysis): ανακάλυψη μοτίβων που περιγράφουν ισχυρώς συσχετισμένα μεταξύ τους γνωρίσματα των δεδομένων.
    - Τα μοτίβα αυτά συνήθως εκφράζονται ως κανόνες πιθανοκρατικής συνεπαγωγής, του τύπου «αν ένα πρότυπο έχει τα γνωρίσματα Α και Β, τότε είναι πολύ πιθανόν να έχει και το γνώρισμα Γ».
  - Δ) **Ανίχνευση ανωμαλιών**: αυτόματος προσδιορισμός των στοιχείων που διαφέρουν σημαντικά από το μεγαλύτερο μέρος των υπολοίπων δεδομένων. Τέτοια πρότυπα ονομάζονται ανωμαλίες (anomalies, outliers).
    - Ένας καλός ανιχνευτής ανωμαλιών πρέπει πράγματι να εντοπίζει τις περισσότερες πραγματικές ανωμαλίες και να δίνει σπάνια ψευδείς συναγερμούς.
    - Χαρακτηριστικά παραδείγματα εφαρμογής είναι η ανίχνευση εισβολέων σε δίκτυα ή η ανίχνευση απάτης (π.χ., σε τραπεζικές συναλλαγές).

Παράδειγμα ταξινόμησης

- Σύνολο δεδομένων Iris.
  - 50 πρότυπα ανά κλάση, 3 κλάσεις (τρία είδη του ανθοφόρου φυτού «ίρις»), 4 γνωρίσματα (μορφολογικές μετρήσεις για τις διαστάσεις των σεπάλων και των πετάλων του κάθε άνθους).
  - Άρα κάθε πρότυπο/φυτό αναπαρίσταται ως ένα 4-διάστατο πραγματικό διάνυσμα.

Παραδείγματα

Iris Versicolor

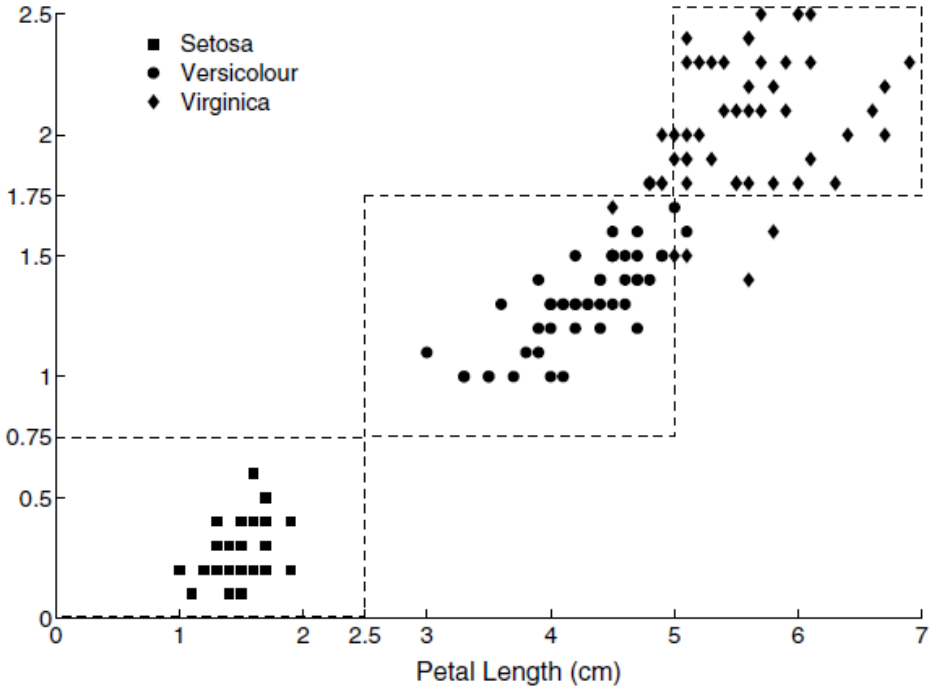
Iris Setosa

Iris Virginica

Παραδείγματα

Παράδειγμα ταξινόμησης

- Έστω ότι διατηρούμε μόνο δύο από τα τέσσερα γνωρίσματα (π.χ., τις μετρήσεις των πετάλων).
  - Μπορούμε πλέον να οπτικοποιήσουμε απευθείας το σύνολο δεδομένων στο ευκλείδειο επίπεδο.



- Άρα μπορούμε να κατασκευάσουμε ένα απλό **σύστημα κανόνων** με δύο κατώφλια (είτε επί του οριζόντιου, είτε επί του κατακόρυφου άξονα) για ταξινόμηση νέων προτύπων.
- Ωστόσο, ένας **ταξινομητής μάθησης** θα αποδώσει καλύτερα.

Πηγή: Tan, “Introduction to Data Mining”, 2006

Παράδειγμα ανάλυσης συσχετίσεων

- Έστω το σύνολο δεδομένων λιανικών πωλήσεων:

Transaction ID	Items
1	{Bread, Butter, Diapers, Milk}
2	{Coffee, Sugar, Cookies, Salmon}
3	{Bread, Butter, Coffee, Diapers, Milk, Eggs}
4	{Bread, Butter, Salmon, Chicken}
5	{Eggs, Bread, Butter}
6	{Salmon, Diapers, Milk}
7	{Bread, Tea, Sugar, Eggs}
8	{Coffee, Sugar, Chicken, Eggs}
9	{Bread, Diapers, Milk, Salt}
10	{Tea, Eggs, Cookies, Diapers, Milk}

Πηγή: Tan, “Introduction to Data Mining”, 2006

Παραδείγματα

- Στόχος ενός αλγορίθμου ανάλυσης συσχετίσεων είναι, π.χ., να εντοπίσει αυτομάτως από αυτά τα δεδομένα ότι ένα καλάθι το οποίο περιέχει πάνες (diapers) αναμένεται με μεγάλη πιθανότητα να περιέχει και γάλα (milk).
  - Εύρεση προϊόντων τα οποία αγοράζονται συνήθως μαζί.

Παραδείγματα

Παράδειγμα ομαδοποίησης

- Έστω το παρακάτω σύνολο αναπαραστάσεων εγγράφων:

Πηγή: Tan, "Introduction to Data Mining", 2006

Article	Words
1	dollar: 1, industry: 4, country: 2, loan: 3, deal: 2, government: 2
2	machinery: 2, labor: 3, market: 4, industry: 2, work: 3, country: 1
3	job: 5, inflation: 3, rise: 2, jobless: 2, market: 3, country: 2, index: 3
4	domestic: 3, forecast: 2, gain: 1, market: 2, sale: 3, price: 2
5	patient: 4, symptom: 2, drug: 3, health: 2, clinic: 2, doctor: 2
6	pharmaceutical: 2, company: 3, drug: 2, vaccine: 1, flu: 3
7	death: 2, cancer: 4, drug: 3, public: 4, health: 3, director: 2
8	medical: 2, cost: 3, increase: 2, patient: 2, health: 3, care: 1

  - Κάθε έγγραφο έχει αναπαρασταθεί ως ένα διάνυσμα η *i*-οστή τιμή του οποίου είναι το πλήθος εμφανίσεων της *i*-οστής λέξης-κλειδιού, από ένα σταθερό προκαθορισμένο λεξιλόγιο λέξεων-κλειδιών.

- Στόχος ενός αλγορίθμου ομαδοποίησης είναι να αναγνωρίσει αυτομάτως δύο διαφορετικές θεματολογίες στο σύνολο των εγγράφων και να αναθέσει καθένα από τα 8 έγγραφα στη μία εκ των δύο.

Thank you for your attention!

Q & A

Contact: imademlis@aueb.gr