ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
Τμήμα Ηλεκτρολόγων & Ηλεκτρονικών Μηχανικών

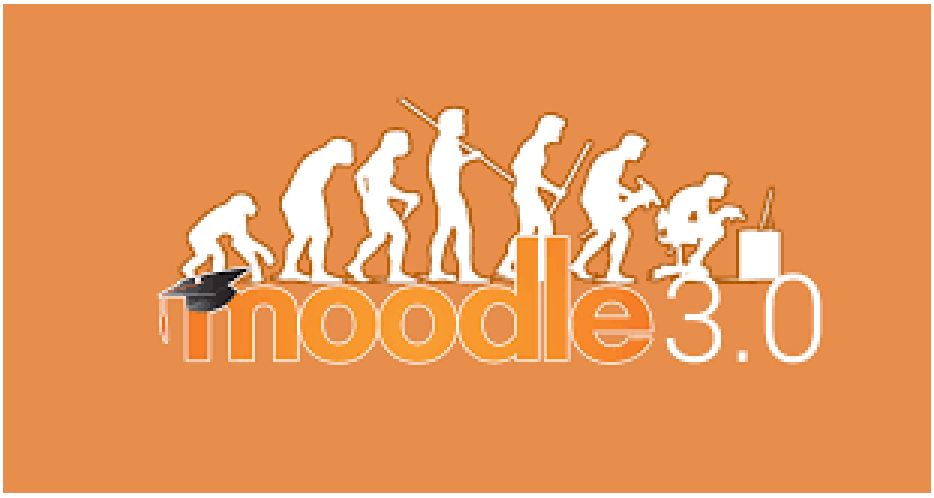
UNIVERSITY of WEST ATTICA
FACULTY OF ENGINEERING
Department of Electrical & Electronics Engineering

Πρόγραμμα Μεταπτυχιακών Σπουδών
Διαδικτυωμένα Ηλεκτρονικά Συστήματα

Master of Science in
Internetworked Electronic Systems

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μέθοδοι οπτικοποίησης δεδομένων από εξόρυξη, με εφαρμογή σε εκπαιδευτικά δεδομένα προερχόμενα από πλατφόρμες ηλεκτρονικής μάθησης



Μεταπτυχιακός Φοιτητής: Ηλίας Μισαηλίδης, Α.Μ.: 0029
Επιβλέπουσα: Μαρία Ραγκούση, Καθηγήτρια

ΑΙΓΑΛΕΩ, ΣΕΠΤΕΜΒΡΙΟΣ 2018

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
Τμήμα Ηλεκτρολόγων & Ηλεκτρονικών Μηχανικών

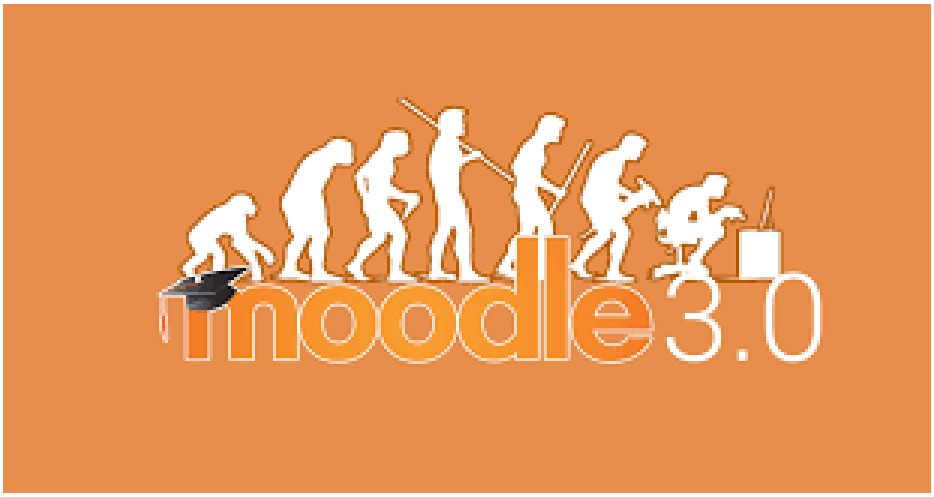
UNIVERSITY of WEST ATTICA
FACULTY OF ENGINEERING
Department of Electrical & Electronics Engineering

Πρόγραμμα Μεταπτυχιακών Σπουδών
Διαδικτυωμένα Ηλεκτρονικά Συστήματα

Master of Science in
Internetworked Electronic Systems

MSc Thesis

Methods for the visualization of mined data, applied in educational data mined from e-learning platforms



Student: Elias Misailidis, Reg. Nr. 0029
MSc Thesis Supervisor: Maria Rangoussi, Professor

ATHENS-EGALEO, SEPTEMBER 2018

ΠΕΡΙΛΗΨΗ

Πολλές και διαφορετικές μέθοδοι έχουν αναπτυχθεί για την αυτοματοποιημένη οπτικοποίηση των δεδομένων που εξορύσσονται από βάσεις δεδομένων ή βάσεις γνώσης. Η παρούσα διπλωματική εργασία προσεγγίζει το πεδίο αυτό με τη μέθοδο top-down (από τον χρήστη προς το σύστημα) ώστε να δώσει εποπτικά την εικόνα των γνώσεων που έχουν συγκεντρωθεί στο πεδίο της εξόρυξης και οπτικοποίησης δεδομένων. Στο πρώτο κεφάλαιο γίνεται ανασκόπηση των μαθησιακών εργαλείων που έχει σήμερα στη διάθεσή του ο χρήστης ενός συστήματος ηλεκτρονικής μάθησης (e-Learning) και γνωριμία με τις σύγχρονες μεθόδους εξόρυξης και οπτικοποίησης δεδομένων. Από το δεύτερο κεφάλαιο και μετά, γίνεται προσπάθεια εμβάθυνσης στα συστήματα αυτά, για την καλύτερη κατανόηση των μεθόδων εξόρυξης και οπτικοποίησης και αναλύονται οι κατάλληλες τεχνικές και εργαλεία για την απόκτηση των επιθυμητών αποτελεσμάτων και στην επιθυμητή μορφή απεικόνισης.

Στο δεύτερο μέρος της διπλωματικής εργασίας παρουσιάζεται αναλυτικά η κωδικοποίηση (προγραμματισμός) των τεχνικών και εργαλείων που έχουν αναφερθεί στο πρώτο μέρος με στόχο την εξαγωγή των επιθυμητών αποτελεσμάτων από την βάση δεδομένων. Κεντρικό ρόλο στο δεύτερο μέρος παίζουν τα εκπαιδευτικά / μαθησιακά δεδομένα που αποθηκεύονται αυτόματα στις βάσεις δεδομένων των πλατφορμών ηλεκτρονικής μάθησης, όπως π.χ. το moodle, καθώς οι εκπαιδευόμενοι αλληλεπιδρούν με το μαθησιακό υλικό.

Η επιλεγμένη μέθοδος αυτόματης οπτικοποίησης κωδικοποιήθηκε ως ένα Plug-in για την πλατφόρμα moodle, με στόχο τη διευκόλυνση του διδάσκοντος, στον οποίο η οπτικοποίηση προσφέρει τα δεδομένα που παράγονται κατά τις συνεδρίες ηλεκτρονικής μάθησης σε μορφή και άποψη εύκολα κατανοητή. Το plug-in εφαρμόστηκε στα εκπαιδευτικά δεδομένα της πλατφόρμας moodle που συντηρεί και χρησιμοποιεί το Τμήμα Ηλεκτρολόγων & Ηλεκτρονικών Μηχανικών του Πανεπιστημίου Δυτικής Αττικής. Στο τελευταίο μέρος της διπλωματικής εργασίας παρουσιάζονται σε μορφή γραφημάτων τα αποτελέσματα από την εφαρμογή του plug-in σε πραγματικά

συλλεγμένα δεδομένα από εκπαιδευτική διαδικασία που λαμβάνει χώρα στον server του Τμήματος κατά τα τρία (3) πλέον πρόσφατα ακαδημαϊκά έτη 2015-16, 2016-17 και 2017-18.

ΛΕΞΕΙΣ – ΚΛΕΙΔΙΑ: Data mining, Data visualization, Learning analytics, E-Learning, moodle, PHP, Database, DBMS.

ABSTRACT

Various methods have been developed for the visualization of data extracted from databases or knowledge bases. This thesis uses a top-down approach to present an overview of data mining and data visualization methods and tools and of the relevant accumulated knowledge.

The first chapter reviews the tools used in e-Learning and offers a brief acquaintance with modern data mining methods. The second and following chapters get into more depth towards the understanding of data mining methods and of their functionalities. The appropriate techniques and tools that yield the sought results are also analyzed.

The second part of this thesis presents in detail the development and coding (programming) of the techniques and tools mentioned in the first part, for data extraction and visualization. The central role in this second part is held by the educational (learning) data extracted from databases where they are automatically being stored while learners interact with e-learning platforms, such as moodle.

The selected visualization method is coded as a plug-in for the moodle platform, aimed to offer the instructors automated and meaningful views or aspects of the educational data produced during e-learning sessions. The plug-in is applied to the educational data stored in the database of the moodle platform used by Department of Electrical and Electronics Engineering, University of West Attica. The last part of this thesis presents in graphical form the results from the application of this plug-in to the real-field data stored in the departmental moodle server during the last three academic years (2015-16, 2016-17, 2017-18).

KEYWORDS: Data mining, Data visualization, Learning analytics, E-Learning, moodle, PHP, Database, DBMS.

ΕΥΧΑΡΙΣΤΙΕΣ

Ξεκινώντας αυτή τη διπλωματική εργασία θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα καθηγήτρια κα. Μαρία Ραγκούση που μου έδωσε την ευκαιρία και την έμπνευση να ασχοληθώ κατ' αρχήν με τις μεταπτυχιακές σπουδές και έπειτα με αυτό το πολύ ενδιαφέρον θέμα σχετικά με την πλατφόρμα e-Learning και τον προγραμματισμό, καθώς και για την καθοδήγηση και την υπομονή που μου έδειξε κατά την εξέλιξη αυτής της διπλωματικής εργασίας. Επίσης θέλω να ευχαριστήσω και όλους αυτούς που με βοήθησαν για να ολοκληρώσω τη παρούσα διπλωματική, είτε δίνοντάς μου χρήσιμες πληροφορίες και γνώσεις είτε βοηθώντας με στα πρακτικά προβλήματα του προγραμματισμού. Η βοήθειά τους αποδείχθηκε αρκετά σημαντική και από πλευράς συλλογής πληροφοριών καθώς και υλοποίησης αυτών που έμαθα στην διάρκεια εκπόνησης αυτής της διπλωματικής. Πάνω απ' όλα όμως θέλω να ευχαριστήσω τον υιό μου που μου υπενθυμίζει το πιο βασικό στη ζωή, κάθε μέρα: ότι ηττημένος δεν είναι αυτός που έχασε μια μάχη άλλα αυτός που δεν προσπάθησε ποτέ να την κερδίσει!

Στον υιό μου, στη γυναίκα μου και στην οικογένεια μου

TABLE OF SYMBOLS – ACRONYMS – ABBREVIATIONS

E-book = Ηλεκτρονικό Βιβλίο

E-Learning = Ηλεκτρονική Μάθηση

E-Class = Ηλεκτρονική Αίθουσα Διδασκαλίας (όνομα γνωστής πλατφόρμας ηλεκτρονικής μάθησης)

Open Source = (Λογισμικό) Ανοικτού Κώδικα

LDAP Server = Lightweight Directory Access Protocol

RTE = Πραγματικός Χρόνος Εκπαίδευσης

DM = Εξόρυξη Δεδομένων (Data Mining)

DA = Ανάλυση Δεδομένων (Data Analytics)

DW = Αποθετήριο Δεδομένων (Data Warehouse)

OLTP = Online Transaction Processing

OLAP = Online Analytical Processing

DB = Βάση Δεδομένων (Database)

DBMS = Σύστημα Διαχείρισης Βάσεων Δεδομένων

Queries = Προγραμματιστικά Ερωτήματα προς τη βάση δεδομένων

AI = Τεχνητή Νοημοσύνη (Artificial Intelligence)

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΗΛΕΚΡΟΝΙΚΗ ΜΑΘΗΣΗ

ΚΕΦΑΛΑΙΟ 2. ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ (DATA MINING)

ΚΕΦΑΛΑΙΟ 3. ΟΠΤΙΚΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ (DATA VISUALIZATION)

ΕΙΣΑΓΩΓΗ

Σε μια κοινωνία της οποίας οι ρυθμοί διαρκώς αυξάνονται και με τον σύγχρονο άνθρωπο να περνά πλέον το μεγαλύτερο χρόνο της ημέρας στην εργασία του και στις υποχρεώσεις του στον υπολογιστή, με την μάθηση λόγω πίεσης χρόνου να συμβαίνει όλο και περισσότερο στην ηλεκτρονική της μορφή (e-Learning), γίνεται επιτακτική η ανάγκη ανάπτυξης εργαλείων και μεθόδων που θα συμπιέζουν τον χρόνο απασχόλησης του χρήστη ή εκπαιδευόμενου με τον Η/Υ και θα ελευθερώνουν χρόνο από τη μάθηση ώστε να αφιερωθεί αλλού. Το πρόβλημα που υπάρχει είναι ότι ο σύγχρονος άνθρωπος θα πρέπει να κρίνει την δουλειά του ή την δουλειά των άλλων γρήγορα, αποτελεσματικά και με σαφή κριτήρια, πράγμα που με την ηλεκτρονική μάθηση (e-Learning) μπορεί να επιτευχθεί. Βέβαια στην περίπτωση της ηλεκτρονικής μάθησης θα πρέπει να υπάρχει πρόνοια ώστε η παροχή των γνώσεων να γίνεται με τρόπο κατανοητό και εύκολο. Αυτός ο στόχος αυτήν την εποχή είναι σε εξέλιξη, με μελλοντική προοπτική πολλών ετών.

Επειδή κάθε ανάγκη δημιουργεί ζήτηση, αναπτύχθηκαν και τα αντίστοιχα εργαλεία λογισμικού για την κάλυψη των αναγκών αυτών και την λύση των προβλημάτων του σύγχρονου ανθρώπου. Το σύνολο των εργαλείων που θα εξεταστούν εδώ ονομάζεται Εξόρυξη Δεδομένων (Data Mining). Το Data Mining είναι ένας κλάδος της Πληροφορικής με «πεδίο δόξης λαμπρό», ταυτόχρονα όμως είναι και ένα εργαλείο που μεσολαβεί ανάμεσα στον χρήστη και την εφαρμογή, άρα ένα εργαλείο που απευθύνεται κυρίως στους ειδικούς του προγραμματισμού Η/Υ. Έτσι δημιουργήθηκε η ανάγκη ο απλός χρήστης να παίρνει την χρήσιμη πληροφορία που προκύπτει από το Data Mining σε μορφή κατανοητή, εύκολα διαχειρίσιμη και πολύ γρήγορα επεξεργάσιμη. Έτσι δημιουργήθηκε η ανάγκη του Data Visualization ή, κοινώς, η οπτικοποίηση των δεδομένων που ο χρήστης θα βλέπει ή θα θέλει να δει, σε μορφή που έχει νόημα και είναι άμεσα κατανοητή από αυτόν. Τα πεδία Data mining, Data analytics και Data visualization συνθέτουν πρακτικά ένα ενιαίο τομέα της Μηχανικής Λογισμικού και της Πληροφορικής γενικότερα, απλά το καθένα αναφέρεται σε διαφορετικά επίπεδα χειρισμού μέχρι ο χρήστης να δει την πληροφορία στην μορφή που επιθυμεί και μπορεί να κατανοήσει.

Η αξία του να δοθούν στον χρήστη εύκολα και κατανοητά δεδομένα στην μορφή που τα επιθυμεί είναι μεγάλη, ιδίως στην «αγορά» της Πληροφορικής, όπου τα εργαλεία που αναζητά ο χρήστης πρέπει να λύνουν τις άμεσα δημιουργούμενες ανάγκες του αλλά και αντίστροφα, στο βαθμό που αυτά τα εργαλεία καλύπτουν τις ανάγκες αυτομάτως γίνονται και ευρέως αποδεκτά από την αγορά.

Η παρούσα εργασία ασχολείται με την πλατφόρμα ηλεκτρονικής μάθησης moodle, μια πλατφόρμα ελεύθερου λογισμικού ανοικτού κώδικα (free & open source) που χρησιμοποιείται ευρέως σήμερα από επιστημονικό προσωπικό, φοιτητές, επαγγελματίες της εκπαίδευσης και της κατάρτισης και άλλες ομάδες με ειδικά ενδιαφέροντα σε όλο τον κόσμο. Στόχος της εργασίας είναι να αναπτύξει πρόσθετα εργαλεία που κάνουν το moodle πιο φιλικό στον απλό χρήστη, οπτικοποιώντας κατάλληλα τις πληροφορίες και τα δεδομένα που τον ενδιαφέρουν και τα οποία θα μπορεί να διαχειριστεί σε χρόνο ταχύτερο και με μεγαλύτερη ευκολία.

1.1. Χρησιμότητα ηλεκτρονικής μάθησης

Η τεχνολογική έκρηξη που παρακολουθούμε όλοι στην εποχή μας δεν θα μπορούσε φυσικά να αφήσει ανεπηρέαστη και την Εκπαίδευση. Σήμερα, με τον τεχνολογικό εξοπλισμό να αυξάνεται ραγδαία, τα δίκτυα των υπολογιστών να έχουν επεκταθεί και ποσοτικά και ποιοτικά σε όλους τους τομείς της ζωής μας και το Διαδίκτυο να κινείται σε ταχύτητες φωτός (με τις οπτικές ίνες), η λύση της παροχής εκπαίδευσης μέσω του υπολογιστή έγινε σε πρώτη φάση μια πολύ ελκυστική υπόθεση και στη συνέχεια μια επιτακτική ανάγκη.

Όπως αναφέρθηκε και στην Εισαγωγή, η έλλειψη χρόνου αλλά και η αδυναμία φυσικής παρουσίας σε πολλές δραστηριότητες στην ζωή του σύγχρονου ανθρώπου δημιούργησε την ανάγκη της διοχέτευσης της πληροφορίας και της γνώσης μέσω του διαδικτύου και των Η/Υ. Έτσι λοιπόν δημιουργήθηκε η μορφή εκπαίδευσης που ονομάζεται ηλεκτρονική μάθηση (e-learning) που είναι ευέλικτη για τον σύγχρονο άνθρωπο, γρήγορη και εύκολα πλέον προσβάσιμη, χάρη στις Τεχνολογίες Πληροφορίας και Επικοινωνίας (ΤΠΕ) που έχουν αναπτυχτεί σήμερα. Όσο κινείται η τεχνολογία με αυτούς τους ρυθμούς, τόσο η εκπαίδευση θα γίνεται πιο ευέλικτη και φυσικά με διακίνηση μεγαλύτερου όγκου πληροφοριών που θα μπορεί να διοχετεύσει. Αυτό κάνει το e-Learning ένα ισχυρό εργαλείο για την εκπαίδευση καθώς ο χρόνος αναζήτησης και εύρεσης οποιασδήποτε υπάρχουσας γνώσης ή απλής πληροφορίας έχει σχεδόν μηδενιστεί.

Όμως τι ακριβώς σημαίνει ο όρος e-Learning? Στην εκπαίδευση, η ηλεκτρονική μάθηση συμβαίνει όταν ένας εκπαιδευτής / καθηγητής / διδάσκων και ένας μαθητής / χρήστης / εκπαιδευόμενος μπορούν απομακρυσμένα, δηλαδή από διαφορετικά σημεία, να επικοινωνούν μεταξύ τους για εκπαιδευτικούς λόγους. Αυτή η σύνδεση μπορεί φυσικά να γίνει με διάφορες τεχνολογίες, όπως με αυτές που προαναφέρθηκαν, αλλά και με την χρήση δορυφορικών τεχνολογιών, την χρήση φορητών συσκευών, Smartphone's, κτλ. **[1], [35]**.

Η ηλεκτρονική μάθηση στηρίζεται σε πλατφόρμες μέσω Η/Υ με κατάλληλο λογισμικό που παρέχουν το εκπαιδευτικό υλικό πολλές φορές και real time (σε πραγματικό χρόνο), ώστε να παρέχεται η μόρφωση και η εκπαίδευση στον χρήστη αδιάκοπα. Η παροχή της γνώσης αυτής της μορφής μπορεί να γίνει, με την σημερινή διαδικτυακή τεχνολογία, και από οποιοδήποτε σημείο και σε οποιοιδήποτε χρόνο. Αυτή η αποδέσμευση από χωρικά ή χρονικά πλαίσια κάνει πολύ φιλικό προς το χρήστη αυτό τον τρόπο μάθησης, ενώ το λογισμικό που είναι επίσης προσαρμοσμένο σε αυτόν τον ρόλο κάνει την μάθηση που προσφέρει πιο ελκυστική και γρήγορη για τον χρήστη.

Δυστυχώς η χρησιμότητα αλλά και οι δυνατότητες της ηλεκτρονικής μάθησης στην Ελλάδα έγιναν αργά αντιληπτές, ενώ αντίθετα σε άλλες χώρες πλέον η εκπαίδευση από μια ηλικία και μετά στηρίζεται όλο και περισσότερο στο e-Learning. Στην χώρα μας γίνεται τα τελευταία χρόνια μια προσπάθεια αξιοποίησης της ηλεκτρονικής μάθησης κυρίως από τις ιδιωτικές επιχειρήσεις, με πρόσβαση στην τεχνολογία αυτή από περισσότερους χρήστες, με τις επενδύσεις των εταιρειών στον τομέα της Πληροφορικής και με το άνοιγμα της τεχνολογίας (σε τεχνικές δυνατότητες λόγω δικτύου και υλικού) σε περισσότερο κόσμο. Φυσικά γίνεται και από τους Δημόσιους φορείς αντίστοιχη προσπάθεια να «τρέξει» αυτού του είδους η εκπαίδευση, με πολλές θετικές προοπτικές στον ορίζοντα.

Μια από τις βασικότερες χρήσεις του e-Learning είναι στην ακαδημαϊκή / πανεπιστημιακή μόρφωση αλλά και στην τεχνολογική εκπαίδευση. Οι επιχειρήσεις Πληροφορικής αποτελούν τον μεγαλύτερο χώρο ανάπτυξης αυτού του είδους μάθησης και εκπαίδευσης παρέχοντας ειδικά σχεδιασμένες πλατφόρμες, όχι τόσο συχνά open source, για την εκπαίδευση που οι ίδιες θέλουν να προσφέρουν. Τα πανεπιστήμια, κυρίως λόγω των πνευματικών δικαιωμάτων που ισχύουν, δεν μπορούν να κατοχυρώσουν εξίσου εύκολα ένα ηλεκτρονικό βιβλίο (e-book) που θα το αναρτούσαν σε μια πλατφόρμα εκπαίδευσης σε σχέση με ένα τυπωμένο σύγγραμμα. Αποτέλεσμα είναι να μην αναπτύσσεται η ηλεκτρονική μάθηση με την ταχύτητα που θα μπορούσε να προσφέρει η τεχνολογία. Λύσεις που έχουν δοθεί στο πρόβλημα αυτό, όπως οι άδειες χρήσης τύπου «Creative Commons», έχουν δώσει μεγάλη ώθηση σε πλατφόρμες ηλεκτρονικής μάθησης όπως οι EdX και Coursera στο εξωτερικό, ή τα «Ανοικτά Ακαδημαϊκά Μαθήματα» στο εσωτερικό. Έχουν όμως τα ελληνικά ΑΕΙ ήδη αναπτύξει σε σχέση με τις ιδιωτικές εταιρείες πολλά εργαλεία λογισμικού της κατηγορίας open source. Τα πανεπιστήμια, λόγω του όγκου του μαθησιακού υλικού αφενός και του πλήθους των φοιτητών τους αφετέρου, θα

μπορούσαν να γίνουν άμεσα οι κύριοι φορείς χρήσης και διάδοσης του e-Learning. Πράγματι, τα πανεπιστήμια που δημιούργησαν τέτοιες πλατφόρμες για e-Learning, τόσο στην Ελλάδα όσο και διεθνώς, θεωρούνται πολύ επιτυχημένα για το επίπεδο εκπαίδευσης που προσφέρουν.

Στα σχολεία, το e-Learning έχει μπει με παιδαγωγικό τρόπο στην ζωή των παιδιών και έχει ξεπεράσει σε πολλές περιπτώσεις τα προβλήματα υλικοτεχνικών υποδομών που παρατηρούνταν τα παλαιοτέρα χρόνια **[1], [22]**.

Οι ιδιωτικές επιχειρήσεις και οι εταιρείες με την ηλεκτρονική μάθηση στοχεύουν κυρίως στην εκπαίδευση του προσωπικού στον χώρο εργασίας τους σαν μια μορφή ευέλικτης μάθησης και ανάπτυξης των γνώσεων και των δεξιοτήτων του προσωπικού. Όμως πολύ σημαντικοί είναι και οι οικονομικοί λόγοι, γιατί προκύπτει εξοικονόμηση δαπανών από μετακινήσεις προσωπικού ή αποφεύγεται η υποχρεωτική απουσία λόγω συμμετοχής σε προγράμματα εκπαίδευσης ή κατάρτισης του προσωπικού μίας επιχείρησης.

1.2. Εκπαίδευση εξ' αποστάσεως

Θα πρέπει πρώτα να οριστεί ο όρος της εκπαίδευσης, πριν τον όρο της εξ' αποστάσεως εκπαίδευσης. Σύμφωνα με τον Ντυρκέμ **[1], [35]**, ένας αρχικός πρωτογενής ορισμός για την εκπαίδευση είναι «*η δράση που κατευθύνεται από τις γενιές των ενηλίκων στις γενιές εκείνες που δεν είναι αρκετά ώριμες για κοινωνική ζωή*». Στην ιστορία της εκπαίδευσης, η εξέλιξη είναι εντυπωσιακή από την αρχαιότητα έως την σημερινή της μορφή που είναι η πιο εξελιγμένη. Πρέπει φυσικά να τονιστεί ότι οι όροι εκπαίδευση, παιδεία και μόρφωση αν και όροι συναφείς δεν αποτελούν ακριβώς συνώνυμα.

Σήμερα ως εκπαίδευση θεωρείται κάθε προσπάθεια απόκτησης γνώσης και δεξιοτήτων με την μορφή των οργανωμένων σε λειτουργικό σύνολο γνώσεων, ικανοτήτων και δεξιοτήτων, οι οποίες αποκτώνται μέσω οργανωμένων προσπαθειών και εργαλείων. Η εκπαίδευση και η μάθηση είναι όροι αλληλένδετοι για αυτό και η εκπαίδευση μπορεί να οριστεί ως οποιαδήποτε μαθησιακή δραστηριότητα που οργανώνεται από κάποιον φορέα, ιδιωτικό ή δημόσιο, με στόχο την βελτίωση των δεξιοτήτων ή των γνώσεων των εμπλεκομένων.

Η εκπαίδευση στην γενική της μορφή μπορεί να χωριστεί σε διάφορες κατηγορίες με πολλά και διαφορετικά κριτήρια. Δύο ενδιαφέρουσες κατηγορίες είναι (α) η δια βίου μάθηση/εκπαίδευση και (β) η εξ αποστάσεως μάθηση/εκπαίδευση.

Η **δια βίου μάθηση/εκπαίδευση** ορίζεται ως ένας κύκλος μάθησης ο οποίος ξεκινά από την παιδική ηλικία με την υποχρεωτική εκπαίδευση (Πρωτοβάθμια, Δευτεροβάθμια) προχωρά ενδεχομένως στην Τριτοβάθμια, ακαδημαϊκή παιδεία και συνεχίζει δια βίου, συμπεριλαμβανομένης και της μάθησης/εκπαίδευσης των ενηλίκων, είτε σε υποχρεωτική (λόγω επαγγελματικών πλαισίων) είτε σε ελεύθερα επιλεγόμενη, ανεξάρτητη μορφή – και όλα αυτά δια βίου **[1], [35]**.

Η **εξ αποστάσεως εκπαίδευση** όπως προαναφέρθηκε δεν απαιτεί ο εκπαιδευτής και ο εκπαιδευόμενος να βρεθούν ταυτόχρονα στον ίδιο χώρο για να επιτευχθεί η εκπαίδευση. Το μαθησιακό υλικό διατίθεται συνήθως ασύγχρονα και ηλεκτρονικά από τον εκπαιδευτή και ο εκπαιδευόμενος το προσπελαύνει από οποιοδήποτε σημείο προτιμά (σπίτι, διακοπές, κλπ.) και στο χρόνο που επιλέγει. Σπανίως η εξ αποστάσεως εκπαίδευση αφορά την υποχρεωτική εκπαίδευση, κατά την οποία είναι υποχρεωτική η φυσική παρουσία των μαθητών και του δασκάλου ταυτόχρονα και στον ίδιο χώρο (σχολείο, τάξη, εκδρομή, εκδήλωση, κλπ.). Εξαίρεση αποτελούν χώρες όπως η Αυστραλία, ή χώρες με συμπλέγματα νησιών, όπου για πρακτικούς λόγους η υποχρεωτική εκπαίδευση συχνά παρέχεται από μακριά με τα εκάστοτε διαθέσιμα μέσα (τηλεόραση, ραδιόφωνο, σήμερα υπολογιστής και διαδίκτυο). Συχνά όμως η εξ αποστάσεως εκπαίδευση εφαρμόζεται κατά την εκπαίδευση ή την ανάπτυξη δεξιοτήτων ενός ενηλίκου, με απομακρυσμένη παρακολούθηση μέσω e-Learning. Αυτή η μορφή της εκπαίδευσης μπορεί να είναι ευέλικτη και ελαστική εξαιτίας και του πλήθους των συμμετεχόντων που μπορούν να παρακολουθήσουν ένα e-class και των μορφών διδασκαλίας και εξέτασης/αξιολόγησης της κατανόησης που έχουν αναπτυχτεί σήμερα. Πλέον μεγάλες εταιρείες πληροφορικής οργανώνουν μαθήματα εξ αποστάσεως (e-class), με χιλιάδες χρήστες on-line και με real time επιδείξεις μεθόδων ή τεχνικών κλπ. Είναι τέτοιες οι δυνατότητες της τεχνολογίας ηλεκτρονικής μάθησης που έχει αναπτυχθεί ώστε να διασφαλίζεται το αδιάβλητο της από απόσταση αξιολόγησης, με αποτέλεσμα σήμερα πολλά τέτοια μαθήματα να χορηγούν πιστοποιητικά. Πράγματι, ο εξ αποστάσεως χρήστης/μαθητής όχι πλέον δεν μπορεί να «κλέψει» κατά την αξιολόγηση/εξέταση με κάποιο τρόπο αλλά σε πειραματική μορφή το λογισμικό μπορεί να

αντιλαμβάνεται και παρακολουθεί και την δυσκολία του κάθε χρήστη στην κατανόηση της κάθε δοκιμασίας ή ερώτησης ενός τεστ, **[1], [35]**.

Η εξ αποστάσεως εκπαίδευση θα μπορούσε επίσης να κατηγοριοποιηθεί και με άλλους τρόπους, όπως για παράδειγμα, με βάση την αλληλεπίδραση που έχει ο διδάσκων με τον χρήστη, διακρίνεται σε **ασύγχρονη αλληλεπίδραση**, **σύγχρονη αλληλεπίδραση** και **εξατομικευμένου ρυθμού (self-paced) αλληλεπίδραση**. Όπως προαναφέρθηκε, η εκπαίδευση και η μάθηση είναι όροι συναφείς. Η μάθηση θα μπορούσε να κατηγοριοποιηθεί επιπλέον σε **ενεργητική μάθηση**, **εποικοδομητική μάθηση** και **εξατομικευμένη μάθηση** όμως η ανάλυσή τους ξεπερνά τον στόχο αυτής της εργασίας, **[1], [35]**.

Η ηλεκτρονική μάθηση είναι σήμερα πλέον η αντιπροσωπευτικότερη μορφή εξ αποστάσεως εκπαίδευσης και παρέχει αμφίδρομη και άμεση επικοινωνία μεταξύ εκπαιδευτή/καθηγητή και εκπαιδευόμενου/φοιτητή. Πλέον υπάρχει διαδραστική σχέση του φοιτητή ή του χρήστη γενικότερα με το υλικό της μάθησης, μέσα από ειδικά εργαλεία, όπως είναι η ζωντανή σύνδεση μέσα στην τάξη με δεκάδες εφαρμογές λογισμικού, είτε με βιντεοσκόπηση είτε με άλλους τρόπους. Η σημαντικότερη μορφή εξ αποστάσεως εκπαίδευσης σήμερα είναι η διαδραστική (interactive learning), όπου ο χρήστης μπορεί σε πραγματικό χρόνο να δώσει π.χ. ένα τεστ δεξιοτήτων σε μια βάση δεδομένων όπου έχουν ήδη εισαχθεί από το διδάσκοντα οι σωστές απαντήσεις, οπότε ο χρήστης παίρνει κατευθείαν ανατροφοδότηση για τις απαντήσεις που έκανε λάθος για διορθώσει το λάθος του ή τροποποιεί τις λάθος αντιλήψεις που είχε διαμορφώσει και προχωρά στην οικοδόμηση νέας, ορθής γνώσης, **[21]**.

Έτσι πλέον η ηλεκτρονική μάθηση είναι αναπόσπαστο κομμάτι της εκπαίδευσης, ειδικά της (επαν)εκπαίδευσης ενός ενηλίκου, αλλά ταυτόχρονα είναι και ένας τομέας με πολύ σημαντικές προεκτάσεις για περαιτέρω ανάλυση και με πολύ σημαντικές προοπτικές για το μέλλον. Κινούμαστε σε ένα χώρο όπου η τεχνολογία ξεπερνά τόσο γρήγορα την ιδία την εκπαιδευτική πράξη, ώστε οι προβλέψεις να δείχνουν ότι σε λίγα χρόνια θα έχει υιοθετηθεί πλήρως και ουσιαστικά σε κάθε εκπαιδευτική βαθμίδα.

1.3. Κατηγορίες e-Learning

Οι κατηγορίες του e-Learning θα μπορούσαν να ομαδοποιηθούν σε πολλές διαφορετικές κατηγορίες, αλλά εδώ επιλέχθηκε για λόγους κατανόησης να χωριστούν ανάλογα με το είδος της γνώσης που προσφέρονται. Έτσι λοιπόν διακρίνονται οι εξής κατηγορίες:

✓ **Τα απλά εργαλεία εκμάθησης και γνώσης**. Πρόκειται για πλατφόρμες κατά κανόνα μεγάλων ιδιωτικών εταιρειών λογισμικού, για την πρόσβαση στις οποίες οι χρήστες δεν χρειάζονται ειδικά προγράμματα και interfaces. Η εταιρεία μέσω του διαδικτύου προσφέρει στο χρήστη πρόσβαση στις βάσεις δεδομένων που τηρεί η ίδια για το κοινό (ουσιαστικά στην ιστοσελίδα της) καθώς και τη δυνατότητα ο χρήστης να διαβάσει on-line το άρθρο ή το βιβλίο ή την παρουσίαση ή το e-book ή την οποιαδήποτε άλλη πηγή πληροφορίας που τον ενδιαφέρει. Πλέον πολλές είναι οι εταιρείες και οι οργανισμοί που παρέχουν μέσω των διαδικτυακών συνδέσμων τους αυτήν την μορφή e-Learning, **[1], [35]**.

✓ **Προσομοιωμένες αίθουσες διδασκαλίας**. Πρόκειται για την πιο εξελιγμένη μορφή του e-Learning που υπάρχει σήμερα και αυτό χάρη στα εργαλεία λογισμικού προσομοίωσης και εκμάθησης που έχουν αναπτυχθεί. Πλέον οι εταιρείες παρέχουν προγράμματα και εργαλεία τα οποία όχι μόνο προσομοιώνουν την εκπαιδευτική αίθουσα, ως προς την μαθησιακή ύλη και την γνώση που προσφέρεται, αλλά παρέχουν πια και αξιόλογη διαδραστικότητα με υπερσύγχρονα εργαλεία (π.χ. live audio και video streaming, multicast) που αντιλαμβάνονται και την δυσκολία στην κατανόηση της ύλης μέσω διαδραστικών τεστ, **[1], [35]**.

✓ **Συστήματα διαχείρισης μάθησης**. Πρόκειται για συστήματα τα οποία διαχειρίζονται εταιρείες και φορείς για την εκπαίδευση, κατάρτιση και επιμόρφωση του προσωπικού τους. Επειδή είναι συστήματα τα οποία διαχειρίζονται οι ίδιες οι εταιρείες, έχουν εξασφαλίσει και τις υλικοτεχνικές υποδομές που η ίδια μπορεί να προσφέρει. Φυσικά είναι μια κατηγορία με τεράστιες προεκτάσεις πλην εκτός του πλαισίου της παρούσας εργασίας. Εντελώς επιγραμματικά θα αναφερθεί ότι τα συστήματα διαχείρισης μάθησης μπορούν να χωριστούν σε συστήματα διαχείρισης μάθησης που απευθύνονται σε πανεπιστημιακούς φορείς και εκπαιδευτικούς οργανισμούς, σε συστήματα διαχείρισης μάθησης για εταιρείες, και σε συστήματα διαχείρισης περιεχομένου μάθησης – μια ιδιαίτερη κατηγορία με πολλές διαφορετικές παραμέτρους, με βασικότερη την

αποθήκευση μεγάλου όγκου μαθησιακού υλικού που έχει συσσωρευτεί στη διάρκεια πολλών ετών και διατίθεται οποτεδήποτε και για οποιαδήποτε αξιοποίηση, **[1], [35]**.

Εικονική εκπαίδευση/μάθηση. Ως εικονική εκπαίδευση (virtual education) εννοείται ένα περιβάλλον σχεδιασμένο από υπολογιστή με ισχυρή αλληλεπίδραση μεταξύ των συμμετεχόντων, οι οποίοι γίνονται συμμέτοχοι σε έναν εικονικό κόσμο (virtual world). Ένα τέτοιο σύστημα θα πρέπει να είναι προσαρμοσμένο στον άνθρωπο και τις ανάγκες του (ανάλογα με την μορφή της εκπαίδευσης) και όχι το αντίθετο. Η δυσκολία που εμφανίζεται εδώ είναι πώς ο χρήστης αλληλεπιδρά με αυτό το εικονικό περιβάλλον. Πρέπει να γίνει αντιληπτό ότι δεν πρόκειται για μια εφαρμογή ή ένα πρόγραμμα αλλά για μια μεθοδολογία που περιγράφεται από ορισμούς και κανόνες. Σε αυτόν τον τομέα συναντώνται οι όροι «τεχνητή ευφυΐα» (artificial intelligence, AI) και «νευρωνικά δίκτυα» (neural networks), **[1], [35]**.

1.4. Πλατφόρμες e-Learning

Η ηλεκτρονική μάθηση υλοποιείται μέσω ειδικών πλατφορμών που προσφέρουν εταιρείες ανάπτυξης λογισμικού ιστοσελίδων κυρίως, οι οποίες μπορούν να δώσουν την δυνατότητα πρόσβασης στο ηλεκτρονικό μαθησιακό υλικό ταυτόχρονα σε πάρα πολλούς χρήστες (σε επίπεδο π.χ. ολόκληρου πανεπιστήμιου ή μεγάλων εταιρειών) και με πάρα πολλές και σύγχρονες δυνατότητες αλληλεπίδρασης με το υλικό και μεταξύ τους. Επιτυγχάνεται έτσι η εξειδικευμένη εκπαίδευση /μάθηση για το κοινό στο οποίο απευθύνονται. Στη συνέχεια θα εξεταστούν οι βασικότερες πλατφόρμες που χρησιμοποιούνται σήμερα, **[35]**.

- ✓ **Η πλατφόρμα moodle**. Το moodle (Modular Object Oriented Developmental Learning Environment) έχει κατά καιρούς χαρακτηριστεί ως ένα open source λογισμικό διαχείρισης μαθημάτων (Course Management System – CMS), ένα σύστημα διαχείρισης μάθησης (Learning Management System – LMS), ένα σύστημα εικονικής μάθησης (Virtual Learning Environment – VLE), ή πιο απλά ένα πακέτο λογισμικού για τη διεξαγωγή ηλεκτρονικών μαθημάτων μέσω Διαδικτύου, που προσφέρει ολοκληρωμένες υπηρεσίες ασύγχρονης τηλεκπαίδευσης. Μπορεί να εγκατασταθεί και σε περιβάλλον Windows και σε Linux/Unix. Δημιουργήθηκε το 1999 από τον Αυστραλό Martin Dougiamas, ως τμήμα του

PhD του, και σύμφωνα με αυτόν έχει δημιουργηθεί πάνω στη γνωστική φιλοσοφία του κοινωνικού εποικοδομητισμού (social constructivism) ενώ το λογότυπο της πλατφόρμας φαίνεται στην επόμενη εικόνα (Εικόνα 1.1). Το moodle παρέχεται δωρεάν ως ελεύθερο λογισμικό ανοικτού κώδικα (κάτω από την GNU Public License) και μπορεί να «τρέξει» σε οποιοδήποτε υπολογιστικό σύστημα που υποστηρίζει τη γλώσσα PHP, ενώ έχει τη δυνατότητα να συνδυάζεται με πολλούς τύπους βάσεων δεδομένων (ιδιαίτερα όμως τη MySQL). Χρησιμοποιείται κυρίως για τις ανάγκες της ασύγχρονης τηλεκπαίδευσης. Μέχρι στιγμής διατίθεται μεταφρασμένο σε περισσότερες από 75 γλώσσες.

Οι ρόλοι που μπορεί να έχει ένας χρήστης της πλατφόρμας moodle είναι βασικά τρεις (διαχειριστής, εκπαιδευτής, εκπαιδευόμενες) με διάφορες παραλλαγές. Οι ρόλοι αυτοί μαζί με τα προνόμια ή δικαιώματα που τους συνοδεύουν, μπορούν να μεταβάλλονται στον βαθμό που ο μαθητής-χρήστης δημιουργώντας και συνεισφέροντας νέο μαθησιακό περιεχόμενο αναβαθμίζεται και ο ίδιος στο πλαίσιο του συστήματος. Αν το moodle εισαχθεί επιτυχώς στην τάξη, η εικονική ανάληψη διαφορετικών ρόλων συνοδεύεται από μια ανάλογη εναλλαγή ρόλων στην πραγματικότητα. Επομένως ο απλός εγγεγραμμένος μαθητής μπορεί να αναβαθμισθεί σε δημιουργό μαθήματος, ή υπεύθυνο για την διδασκαλία μιας ενότητας, ενώ ο δάσκαλος μπορεί να είναι ταυτόχρονα και διαχειριστής ή και μαθητής στο ίδιο ή σε άλλο μάθημα, **[18], [35]**.

Εικόνα.1.1: Logo της Πλατφόρμας Ηλεκτρονικής Μάθησης moodle.
(Πηγή. https://moodle.org/logo/)

✓ **Η πλατφόρμα Open e-Class**. Το Open e-Class είναι ελεύθερο λογισμικό διαχείρισης εκπαιδευτικού περιεχομένου (Course Management System) και αποτελεί εξέλιξη του Open Source “Classroom Online” της Claroline, με πιστοποίηση των χρηστών και των δικαιωμάτων τους μέσω του κεντρικού LDAP Server. Σχεδιάστηκε για συνεργασία και με άλλες πλατφόρμες αλλά και με άλλες διαδικτυακές εφαρμογές όπως email, ενώ το

λογότυπο της φαίνεται στην συνέχεια (Εικόνα 1.2). Μπορεί να εγκατασταθεί επίσης και στα δυο βασικά λειτουργικά συστήματα (Windows και Linux/Unix), ενώ έχει minimum απαιτήσεις εγκατάστασης. Η βάση σε αυτήν την πλατφόρμα είναι το λογισμικό Apache 1.3. Είναι μια αρκετά εύκολη στην εξοικείωση και τη χρήση πλατφόρμα τηλεκπαίδευσης, πολύ φιλική στους χρήστες και με εύκολη περιήγηση. Δημιουργήθηκε από την Ομάδα Ασύγχρονης Τηλεκπαίδευσης του Ελληνικού Ακαδημαϊκού Διαδικτύου GUnet και έχει εγκατασταθεί και χρησιμοποιείται από πολλά ελληνικά ακαδημαϊκά ιδρύματα. Η πρώτη έκδοση του Open e-Class κυκλοφόρησε το Μάρτιο 2003, **[35]**.

Εικόνα.1.2: Logo της Πλατφόρμας Ηλεκτρονικής Μάθησης Open e Class.

(Πηγή. https://ocp.teiath.gr/index.php?localize=el)

✓ **Η πλατφόρμα eFront**. Το eFront είναι μια open source πλατφόρμα τηλεκπαίδευσης. Το eFront έχει σχεδιαστεί ώστε να επιτρέπει την δημιουργία δικτυακών κοινοτήτων μάθησης με σημαντικές ευκαιρίες για συνεργασία και αλληλεπίδραση με λογότυπο που παρουσιάζεται στην επόμενη εικόνα (Εικόνα 1.3). Το σύστημα χρησιμοποιεί μια διεπαφή χρήστη στηριγμένη σε εικονίδια. Η πλατφόρμα προσφέρει ένα πλήθος από χαρακτηριστικά, όπως εργαλεία για την ανάπτυξη (συγγραφή) μαθησιακού περιεχομένου και ανάπτυξη δραστηριοτήτων αξιολογήσεων, ένα εργαλείο διαχείρισης παραδοτέων, ένα πλήθος από στατιστικά στοιχεία, εσωτερικό σύστημα επικοινωνίας, φόρουμ, συνομιλία, δημοσκοπήσεις και άλλα. Το eFront είναι συμβατό και πιστοποιημένο σύμφωνα με το πρότυπο SCORM 1.2 και SCORM 2004 / 4^{η} έκδοση. Πολλά χαρακτηριστικά της πλατφόρμας (π.χ. διαχείριση δεξιοτήτων, οργανόγραμμα, διαχείριση τομέων) την καθιστούν κατάλληλη για χρήση από οργανισμούς και εταιρίες, ιδιαίτερα από τμήματα διαχείρισης ανθρώπινων πόρων. Το eFront είναι μια πολυγλωσσική πλατφόρμα που προσφέρεται σε 40 γλώσσες, **[35]**.

Εικόνα.1.3: Logo της Πλατφόρμας Ηλεκτρονικής Μάθησης efront.

(Πηγή. https://www.efrontlearning.com/pricing)

✓ **Η πλατφόρμα ILIAS**. Το ILIAS είναι ένα σύστημα ασύγχρονης τηλεκπαίδευσης που δημιουργήθηκε από το Πανεπιστήμιο της Κολωνίας (Γερμανία) και το τμήμα έρευνας και επιστημών του Πανεπιστημίου North Rhine (Γερμανία). Χωρίζεται σε περιβάλλον εργασίας για τους μαθητές, όπου παρέχει μια κύρια γραμμή εργασιών με διάφορες λειτουργίες και σε περιβάλλον εργασίας για τους διδάσκοντες. Το σύστημα τηλεκπαίδευσης ILIAS που το λογότυπο της το βλέπουμε στην επόμενη εικόνα (Εικόνα 1.4) μπορεί να εγκατασταθεί και σε περιβάλλον Windows και σε Linux/Unix, **[35]**.

Εικόνα.1.4: Logo της Πλατφόρμας Ηλεκτρονικής Μάθησης ILIAS.

(Πηγή. https://commons.wikimedia.org/wiki/File:Logo_ILIAS.svg)

✓ **Η πλατφόρμα ELEDGE**. Το ELEDGE δημιουργήθηκε από το Πανεπιστήμιο της Utah (ΗΠΑ) με σκοπό την διδασκαλία μαθημάτων διαδικτυακά. Είναι μια συλλογή servers με βάση τη βάση δεδομένων MySQL. Σαν όλες τις πλατφόρμες τηλεκπαίδευσης, έχει πολλές από τις δυνατότητες ηλεκτρονικής μάθησης όπως κουΐζ, τεστ, εργασίες κτλ. και μπορεί να εγκατασταθεί και σε Windows και σε Linux/Unix. Το λογότυπο της φαίνεται στην συνέχεια (Εικόνα 1.5), **[35]**.

Εικόνα.1.5: Logo της Πλατφόρμας Ηλεκτρονικής Μάθησης ELEDGE.

(Πηγή. http://www.ranklogos.com/logos/colleges-and-universities-logos/page/23/)

✓ **Η πλατφόρμα CLAROLINE.** Το CLAROLINE σχεδιάστηκε από το Πανεπιστήμιο UCL (University of Louvain) του Βελγίου με λογότυπο όπως εμφανίζεται στην συνέχεια (Εικόνα 1.6) και ξεπερνά σήμερα τα 400 μαθήματα e-Learning που προσφέρονται στο Πανεπιστήμιο αυτό. Την πλατφόρμα αυτή την έχουν υιοθετήσει πάνω από 250 πανεπιστημιακά ιδρύματα στον κόσμο για μαθήματα e-Learning. Σαν πλατφόρμα είναι απλή, χωρίζεται σε διαφόρους τομείς που αφορούν διαφορετικά τον κάθε εγγεγραμμένο (φοιτητής, εκπαιδευτικός, απλός επισκέπτης για courses, κτλ.) ενώ για τον καθηγητή υπάρχουν ειδικά εργαλεία παρακολούθησης την ενεργειών των φοιτητών. Το CLAROLINE υποστηρίζεται και αυτό και σε Windows και σε Linux/Unix. Δεν έχει δικό του text editor και η ανταλλαγή μηνυμάτων γίνεται μέσω e-mail, ενώ κάθε φοιτητής μπορεί να επιτρέψει στον συμφοιτητή του να έχει πρόσβαση στις σημειώσεις ή τα αρχεία του. Τέλος υπάρχουν ανακοινώσεις για το κάθε μάθημα ξεχωριστά, **[35]**.

Εικόνα.1.6: Logo της Πλατφόρμας Ηλεκτρονικής Μάθησης Claroline.

(Πηγή. https://www.slideshare.net/ptsavdar/claroline-connect-edmedia-conference-2016)

1.5. Ευρωπαϊκή και Ελληνική πολιτική e-Learning

Παρόλο που στην Ευρώπη η τηλεκπαίδευση είναι σαφώς πιο ανεπτυγμένη από ότι στην Ελλάδα και πολλές ευρωπαϊκές χώρες την έχουν εντάξει στο εκπαιδευτικό τους σύστημα (κυρίως οι βόρειες χώρες, που έχουν πρότυπα εκπαιδευτικά και κοινωνικά συστήματα) παρόλα αυτά δεν είναι τόσο ανεπτυγμένη όσο στην Αμερική όπου το e-Learning αποτελεί πλέον κύριο κομμάτι κάθε τύπου εκπαίδευσης – ιδίως μετά το λύκειο. Η προσπάθεια της Ευρωπαϊκής Ένωσης να βρει κοινά σημεία και στόχους για την τηλεκπαίδευση στην Ευρώπη σκοντάφτει στις διάφορες γραφειοκρατικές διαδικασίες για την έγκριση και εισαγωγή νέων συστημάτων ή μεθόδων στη δημόσια υποχρεωτική εκπαίδευση και πολλές φορές στις διαφορετικές κουλτούρες των μελών της.

Στο πλαίσιο της κοινής ευρωπαϊκής πολιτικής, η Ένωση έχει εκδώσει τα κείμενα «Ευρώπη της γνώσης» και το «Ο Ρόλος των πανεπιστημίων στην Ευρώπη της γνώσης» που αποτελούν έναν οδηγό για το πώς οραματίζεται η Ευρώπη την εκπαίδευση και τον ρόλο των Πανεπιστημίων σε αυτήν. Σε αυτή την προσπάθεια έχουν εκδοθεί και άλλα πολλά άρθρα σχετικά με την γλώσσα στην εκπαίδευση, όπως το «Κοινό Ευρωπαϊκό Πλαίσιο Αναφοράς για την Γλώσσα» και το «Ευρωπαϊκό Σχέδιο Δράσης E-Learning», **[21], [23]**.

Το μεγαλύτερο άλμα όμως έγινε μετά την Σύνοδο Κορυφής το 2000 όπου οι αρχηγοί των κρατών-μελών αποφάσισαν για μια κοινή, ανταγωνιστική και δυναμική ευρωπαϊκή εκπαίδευση η οποία θα περιελάμβανε αρκετούς τομείς όπως:

- ✓ Τη διδασκαλία σε όλα τα σχολεία όλων των βαθμίδων της χρήσης Η/Υ και την πρόσβαση στο διαδίκτυο όλων των σχολείων των κρατών-μελών.
- ✓ Τη δημιουργία ενός κοινού ευρωπαϊκού δικτύου για την επικοινωνία των ευρωπαϊκών πανεπιστημίων μεταξύ τους και μεταξύ των βιβλιοθηκών τους.
- ✓ Ορίστηκαν ως τόποι μάθησης που θα πρέπει να εξοπλιστούν και αυτά με πρόσβαση στο διαδίκτυο και τα πολιτιστικά κέντρα και τα μουσεία, ώστε να μπορούν να χρησιμοποιούν τις ανοικτές βιβλιοθήκες τους τα σχολεία.
- ✓ Τον εξοπλισμό των σχολείων με το κατάλληλο Hardware και Software και την ανάπτυξη υπηρεσιών λογισμικού για την ενημέρωση και των εκπαιδευτικών σε αυτές τις τεχνολογίες.

Η προσπάθεια της Ένωσης για τον «Ψηφιακό αλφαβητισμό» και την δικτυακή ενοποίηση των βιβλιοθηκών μεγάλων Πανεπιστημίων ήταν πολύ σημαντική και υποστηρίχθηκε μέσα από τα πολλά κοινοτικά προγράμματα και με μεγάλα κονδύλια. Η προσπάθεια είχε ως στόχο ευρωπαίους πολίτες και ιδρύματα που δεν μπορούσαν να έχουν πρόσβαση στις νέες τεχνολογίες ή στην εκπαίδευση/εξειδίκευση. Αυτή η προσπάθεια υλοποιήθηκε με τις συμφωνίες της Ένωσης με τα ευρωπαϊκά πανεπιστήμια για on-line μαθήματα και κατάρτιση που θα προσέφεραν τα ίδια μέσω του διαδικτύου, **[21], [23]**.

Γενικότερα η Ευρωπαϊκή Ένωση προσαρμόζει τα προγράμματά της ανά τριετία ή τετραετία, σε μια προσπάθεια να ακολουθεί πάντα τις τεχνολογικές εξελίξεις και τάσεις.

Οι πρώτες προσπάθειες της Ελλάδας για την αξιοποίηση αυτών των ευκαιριών της Ευρωπαϊκής Ένωσης για την τηλεκπαίδευση και γενικότερα τις ασύγχρονες μορφές εκπαίδευσης ξεκίνησε με το Γ' Κοινοτικό Πλαίσιο Στήριξης και το πρόγραμμα SOCRATES. Το πρόγραμμα αυτό αφορούσε κυρίως και τις προδιαγραφές και την υλοποίηση ενός συστήματος e-Learning και την εγκατάσταση του απαραίτητου εξοπλισμού (Hardware και Software) για την επιμόρφωση και ενημέρωση του προσωπικού και την εξοικείωσή του με αυτές τις τεχνολογίες. Σκοπός ήταν και η δημιουργία ενός μεγάλου ψηφιακού περιβάλλοντος και των υποδομών αυτού.

Στη συνέχεια, το Υπουργείο Παιδείας μέσα από διάφορα Επιχειρησιακά Προγράμματα συμμετείχε σε αυτές τις δράσεις της Ένωσης και έγινε προσπάθεια εξοπλισμού (υλικοτεχνικού) των πανεπιστημιακών ιδρυμάτων της χώρας με στόχο τους τομείς της Παιδείας, του Πολιτισμού, των Επικοινωνιών, της βελτίωσης της ζωής του πολίτη και της μείωσης της γραφειοκρατίας, την εξοικείωση με την τεχνολογία κτλ. για την κάλυψη των εκπαιδευτικών απαιτήσεων της τριτοβάθμιας εκπαίδευσης. Έτσι λοιπόν σήμερα πολλά πανεπιστήμια έχουν αναπτύξει ένα μεγάλο δίκτυο τηλεκπαίδευσης όπου τα μαθήματα και η ύλη προσφέρονται μέσω του διαδικτύου με πολύ σύγχρονες υποδομές και τεχνικές όπως Live Video and Audio Streaming κτλ. Μερικά τέτοια ιδρύματα είναι το Εθνικό Μετσόβιο Πολυτεχνείο, το Καποδιστριακό Πανεπιστήμιο Αθηνών, το Οικονομικό Πανεπιστήμιο Αθηνών, το Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, το Πανεπιστήμιο Κρήτης, Αιγαίου, κ.α., **[21], [23]**.

Στον ιδιωτικό τομέα οι εξελίξεις είναι σαφώς ταχύτερες καθώς πολύ μεγάλες εταιρείες πληροφορικής δίνουν πολύ οικονομικά πακέτα με πολλές διευκολύνσεις ενώ οι

ίδιες οι εταιρείες, έχοντας αναλύσει τα ωφέλει τις τηλεκπαίδευσης για το προσωπικό τους, έχουν κάνει μια μεγάλη στροφή προς την τάση αυτή. Οι ιδιωτικές εταιρείες έχουν πάει ένα βήμα παρακάτω στην τεχνολογία, χρησιμοποιώντας είτε διαδραστική μάθηση είτε με την δημιουργία εικονικών αιθουσών και τρόπων διδασκαλίας επενδύοντας οι ίδιες σε αυτό.

1.6. Μειονεκτήματα και Πλεονεκτήματα e-Learning

Όπως κάθε τεχνολογία που αναπτύσσεται και εφαρμόζεται στους διάφορους τομείς της κοινωνίας, και ειδικά σε έναν τόσο σημαντικό τομέα όπως η εκπαίδευση, έτσι και η ηλεκτρονική μάθηση έχει και πλεονεκτήματα και μειονεκτήματα. Μετά την ανάλυση που προηγήθηκε, θα μπορούσαν να συνοψιστούν τα πλεονεκτήματα και τα μειονεκτήματα του e-Learning ως εξής:

- ✓ Είναι άμεσα και πάντα διαθέσιμο όλο το 24-ωρο, κάθε μέρα, όλο το χρόνο.
- ✓ Δεν υπάρχουν περιορισμοί στην ώρα και στην χώρα από όπου μπορεί να εκπαιδευτεί ο χρήστης.
- ✓ Είναι ιδιαιτέρως ακριβές και αποτελεσματικό αφού υλοποιείται με τα πιο σύγχρονα μέσα πληροφορικής και δεν αφήνει κανένα ίχνος ασάφειας (live Audio &Video streamingκτλ.).
- ✓ Μπορεί να γίνει με διάφορες μεθόδους (ασύγχρονη ή σύγχρονη εκπαίδευση) και ανάλογα τις απαιτήσεις του καθηγητή ή του φοιτητή.
- ✓ Ο εκπαιδευόμενος μπορεί να επικοινωνήσει άμεσα με τον καθηγητή ή με κάποιον συμφοιτητή του.
- ✓ Είναι ευέλικτη μέθοδος ως προς τις υποχρεώσεις των χρηστών, δηλαδή ένας εργαζόμενος μπορεί να παρακολουθήσει ένα μάθημα από το χώρο της εργασίας του, ή ο μαθητής και ο φοιτητής από το σπίτι τους ενώ μπορεί να το μεταθέσει για οποιαδήποτε μελλοντική στιγμή, ανάλογα με τις επείγουσες υποχρεώσεις.
- ✓ Το μαθησιακό υλικό που αναρτάται συνεχώς βελτιώνεται και ανανεώνεται από τον διδάσκοντα και επομένως φτάνουν οι πιο σύγχρονες και επίκαιρες μέθοδοι και γνώσεις άμεσα στον χρήστη.
- ✓ Τόσο ο καθηγητής όσο και ο φοιτητής δεν είναι αδρανείς πομπός και δέκτες, αντίστοιχα, γνώσεων και πληροφοριών, αφού συμμετέχουν άμεσα και σε πραγματικό χρόνο στο μάθημα.

✓ Ο φοιτητής μπορεί να προσαρμόσει την ύλη και το μάθημα ανάλογα με τον δικό του ατομικό τρόπο με τον οποίο μπορεί να μάθει ή να δεχτεί την πληροφορία καλύτερα.

✓ Γίνεται καλύτερη δομή και οργάνωση και της ίδιας της ύλης του μαθήματος αλλά και του χρονικού προγράμματος με βάση το οποίο θα διδαχτούν το κάθε αντικείμενο οι χρήστες.

✓ Ο καθηγητής μπορεί να ελέγξει την πρόοδο του κάθε φοιτητή με εύκολο τρόπο, αφού έχει πλέον στα χέρια του ισχυρά εργαλεία όπως π.χ. τη στατιστική παρακολούθηση της εξέλιξης ενός χρήστη μέσα από τα τεστ αξιολόγησης δεξιοτήτων και γνώσεων της πλατφόρμας, αλλά και μπορεί και ο ίδιος να βελτιωθεί αφού μπορεί να δει που οι φοιτητές/χρήστες υστερούν και τι ακριβώς δεν κατανοούν.

✓ Υπάρχει συνεχής εξέλιξη του τομέα αυτού και των εργαλείων του αφού έχουμε συνεχή άλματα της τεχνολογίας και του software και συνεχώς ανανεωμένες ιδέες.

✓ Ο αριθμός των χρηστών που μπορούν να παρακολουθήσουν με τον τρόπο αυτό ένα διδακτικό αντικείμενο μπορεί να ξεπεράσει κατά πολύ τον αριθμό των χρηστών που θα μπορούσαν με φυσική παρουσία να παρευρεθούν σε έναν χώρο.

✓ Το κόστος του φορέα που διοργανώνει και υλοποιεί ένα εκπαιδευτικό πρόγραμμα με ηλεκτρονικό τρόπο είναι πολύ μικρό αφού ξεπερνιέται η ανάγκη για έντυπα και οι πληροφορίες δίνονται ψηφιακά (π.χ. με e-books ή με σημειώσεις και παρουσιάσεις που οι χρήστες λαμβάνουν ψηφιακά και διαδικτυακά). Αυτό έχει θετική επίπτωση και στο περιβάλλον (με την μειωμένη κατανάλωση χαρτιού, χρήση εκτυπωτών και μελανιού κλπ.)

✓ Ειδικές κατηγορίες της κοινωνίας, όπως τα ΑμΕΑ, μπορούν να έχουν ισότιμη πρόσβαση πλέον στη γνώση και την εκπαίδευση από τον χώρο τους, χωρίς ταλαιπωρία, καλύπτοντας και τις περιπτώσεις που η φυσική τους προσέγγιση και παρουσία στην αίθουσα δεν έχει προβλεφτεί (δυστυχώς) από την πολιτεία ή τον εκπαιδευτικό οργανισμό.

✓ Αυτή η μορφή εκπαίδευσης μπορεί να χρησιμοποιηθεί από διάφορους φορείς, κυρίως ιδιωτικούς, όχι μόνον για εκπαίδευση αλλά και για την κατάρτιση του προσωπικού τους σε συγκεκριμένες δεξιότητες.

✓ Οι μαθητές και ο δάσκαλος χωρίζονται σε ομάδες, ανταλλάσσουν μέσα από την πλατφόρμα μηνύματα, μοιράζονται πόρους, ή -ακόμη- συνδιαμορφώνουν περιεχόμενο μέσα από τα ενσωματωμένα wikis. Στην συνέχεια, καλούν άλλες τάξεις του σχολείου τους (ή και άλλου σχολείου) να μοιραστούν την εργασία τους και να συμμετάσχουν στην έρευνά τους (συνεργατική μάθηση – collaborative learning), **[1], [35], [22], [23]**.

Αντίστοιχα μειονεκτήματα που έχουν παρατηρηθεί είναι:

✓ Δυσκολία χειρισμού των μέσων και των τεχνολογιών των συστημάτων e-Learning από συγκεκριμένες ομάδες χρηστών, είτε λόγω ηλικίας είτε λόγω του ότι δεν έχουν έρθει σε επαφή ποτέ με τέτοια εργαλεία (Η/Υ, Διαδίκτυο κτλ.) είτε λόγω αδυναμίας κατανόησης της γλώσσας που χρησιμοποιούν.

✓ Άρνηση να δεχθούν την νέα μορφή εκπαίδευσης και προσκόλληση στους παλιούς τρόπους μόρφωσης και εκπαίδευσης.

✓ Υπάρχει ο κίνδυνος ο φοιτητής να μην συμμετέχει ουσιαστικά αλλά μόνο τυπικά στην ηλεκτρονική εκπαιδευτική διαδικασία.

✓ Επειδή πολλές φορές οι φοιτητές ρωτούν μεταξύ τους τις απορίες και μοιράζονται μεταξύ τους τις γνώσεις, μπορεί ο διδάσκων να αντιληφθεί κάποια λάθη, παρανοήσεις ή κάποια δυσνόητη έννοια και να συνεχίσει παραβλέποντάς τα, με αποτέλεσμα η υπόλοιπη εκπαιδευτική διαδικασία να εξελίσσεται αρνητικά.

✓ Σε μια πιο φιλοσοφημένη προσέγγιση, μπορεί να διατυπωθεί ο προβληματισμός ότι η τάξη χάνει την έννοια που είχε κάποτε, της φυσικής ανθρώπινης αλληλεπίδρασης και επαφής, αφού όλα γίνονται πλέον μέσω υπολογιστών και οι άνθρωποι αποξενώνονται κοινωνικά. Χάνοντας την επαφή με τους μαθητές του, ο διδάσκων είναι επικίνδυνο σταδιακά να χάσει την ικανότητα μετάδοσης της γνώσης στην αίθουσα.

✓ Αυξάνονται οι οικονομικές απαιτήσεις καθότι για να μπορείς να συνεχίσει ένας φορέας να προσφέρει τις πιο σύγχρονες μεθόδους e-Learning απαιτείται η συχνή επικαιροποίηση / αλλαγή και του Software και Hardware του συνόλου του εξοπλισμού του, που σημαίνει μεγάλο κόστος επένδυσης.

✓ Υπάρχουν επιστήμες και τομείς που το e-Learning δεν θεωρείται ακόμη σήμερα το πλέον αποτελεσματικό εργαλείο, όπως π.χ. στην Ιατρική και αλλού.

✓ Οι μαθητές ενδέχεται να μην μπορούν να κατευθύνουν τον τρόπο σκέψης και μόρφωσής τους, καθώς δεν έχουν την επαφή με τον καθηγητή για να τους καθοδηγήσει στον τρόπο σκέψης και οργάνωσης.

Δημιουργούνται γενιές με μειωμένη ικανότητα έκφρασης στον γραπτό και προφορικό λόγο, αφού μέσω των Η/Υ τα κείμενα ορθογραφικά και συντακτικά δημιουργούνται και διορθώνονται αυτόματα και πολλές φορές δεν απαιτείται ούτε και αυτό από ένα μάθημα e-Learning, **[1], [35], [22], [23]**.

Εξόρυξη Δεδομένων (Data Mining)

2

2.1. Γιατί το Data Mining;

Σε έναν κόσμο όπου καθημερινά συλλέγονται και ανταλλάσσονται τεράστιοι όγκοι δεδομένων, η ανάλυση όλων αυτών των δεδομένων για την εξαγωγή της χρήσιμης πληροφορίας είναι μια σημαντική ανάγκη για τον άνθρωπο. Στο υποκεφάλαιο αυτό εξετάζεται πώς το Data Mining (εξόρυξη δεδομένων) μπορεί να καλύψει αυτή την ανάγκη, παρέχοντας εργαλεία για την εξόρυξη και εν τέλει εξαγωγή ωφέλιμων γνώσεων από πλήθος δεδομένων που οπωσδήποτε περιέχουν πολλή πλεοναστική ή και αντιφατική πληροφορία. Το πεδίο της Εξόρυξης Δεδομένων μπορεί να θεωρηθεί ως εξέλιξη της τεχνολογίας επεξεργασίας των πληροφοριών και της μηχανικής λογισμικού.

Ένα ερώτημα που θα μπορούσε να τεθεί είναι αν ζούμε στην εποχή των πληροφοριών ή στην εποχή των δεδομένων. Μερικοί απαντούν με βεβαιότητα ότι ζούμε στην εποχή των δεδομένων. Terabytes ή Petabytes δεδομένων μεταφέρονται καθημερινά πάνω από τα δίκτυα υπολογιστών, ανταλλάσσονται μεταξύ των χρηστών στο διαδίκτυο, μεταδίδονται ή αποθηκεύονται σε διάφορες συσκευές αποθήκευσης δεδομένων από μία επιχείρηση, από την κοινωνία, τους επιστήμονες, τους μηχανικούς, τον κόσμο της ιατρικής και της υγείας και σχεδόν κάθε πτυχή της καθημερινής μας ζωής. Αυτή η έκρηξη του διαθέσιμου όγκου δεδομένων είναι αποτέλεσμα της ψηφιοποίησης και της ψηφιακής διασύνδεσης της κοινωνίας αλλά και της ταχείας ανάπτυξης ισχυρών εργαλείων συλλογής και αποθήκευσης δεδομένων. Οι επιχειρήσεις σε όλο τον κόσμο δημιουργούν γιγαντιαία σύνολα δεδομένων, συμπεριλαμβανομένων των συναλλαγών των αγορών και πωλήσεων, περιγραφών προϊόντων, προωθήσεων πωλήσεων καθώς και προφίλ εταιρειών και πελατών. Οι επιστημονικές και τεχνικές πρακτικές δημιουργούν συνεχώς υψηλές απαιτήσεις και όγκο δεδομένων της τάξης των Petabytes, προερχόμενων από την τηλεπισκόπηση, τη μέτρηση διαδικασιών, τα επιστημονικά πειράματα, την απόδοση κάποιου συστήματος, τις παρατηρήσεις μηχανικής και πολλά άλλα.

Τα παγκόσμια τηλεπικοινωνιακά δίκτυα μεταφέρουν καθημερινά δεκάδες Petabytes δεδομένων. Η ιατρική και η βιομηχανία υγείας παράγουν τεράστια ποσά δεδομένων από

ιατρικά αρχεία, για την παρακολούθηση ασθενών και για την ιατρική απεικόνιση. Δισεκατομμύρια αναζητήσεις στον Παγκόσμιο Ιστό (World-Wide Web, www) που υποστηρίζονται από μηχανές αναζήτησης (search engines) επεξεργάζονται καθημερινά δεκάδες Petabytes δεδομένων. Οι κοινότητες και τα κοινωνικά μέσα έχουν γίνει όλο και πιο σημαντικές πηγές δεδομένων, κυρίως σε ψηφιακές φωτογραφίες και βίντεο, blogs, κοινότητες του Ιστού και διάφορα είδη κοινωνικών δικτύων. Ο κατάλογος των πηγών που παράγουν τεράστια ποσά δεδομένων είναι ατελείωτος.

Αυτό το ανεξέλεγκτα αυξανόμενο, ευρέως διαθέσιμο και γιγαντιαίο σύνολο δεδομένων καθιστά την εποχή μας πραγματικά εποχή των δεδομένων. Απαιτούνται ισχυρά και ευέλικτα εργαλεία για την αυτόματη αναζήτηση και εύρεση πολύτιμων πληροφοριών από τα τεράστια ποσά δεδομένων και τη μετατροπή αυτών των δεδομένων σε οργανωμένη γνώση. Αυτή η αναγκαιότητα έχει οδηγήσει στη γέννηση της επιστημονικής περιοχής της εξόρυξης δεδομένων (Data Mining). Το πεδίο είναι νέο, δυναμικό και ελπιδοφόρο. Το Data Mining έχει κάνει και θα συνεχίσει να κάνει σπουδαία βήματα στο μέλλον και στην ανερχόμενη επιστήμη των πληροφοριών.

Το Data Mining όπως φαίνεται και από τους τομείς που προκύπτει (Εικόνα 2.1) στοχεύει να μετατρέψει μια μεγάλη συλλογή δεδομένων σε γνώση. Μια μηχανή αναζήτησης (π.χ. Google) λαμβάνει εκατοντάδες εκατομμύρια ερωτημάτων κάθε μέρα. Κάθε ερώτημα μπορεί να θεωρηθεί ως συναλλαγή όπου ο χρήστης περιγράφει την ανάγκη πληροφόρησής του. Είναι ενδιαφέρον ότι ορισμένοι τρόποι αναζήτησης απαντήσεων στα ερωτήματα των χρηστών μπορούν να αποκαλύψουν εξαιρετικά ενδιαφέρουσες ή σημαντικές πληροφορίες που δεν θα μπορούσαν να αποκτηθούν με την απλή ανάκτηση / ανάγνωση των μεμονωμένων στοιχείων των δεδομένων. Για παράδειγμα στην περίπτωση της Γρίπης, κατά την αναζήτηση μέσω της μηχανής Google διαπιστώθηκε στενή σχέση μεταξύ του αριθμού των ατόμων που αναζητούν πληροφορίες σχετιζόμενες με τη γρίπη και του αριθμού των ατόμων που έχουν πραγματικά συμπτώματα γρίπης. Ένα «μοτίβο» (pattern) εμφανίζεται όταν συγκεντρώνονται όλα τα ερωτήματα αναζήτησης που σχετίζονται π.χ. με τη γρίπη. Χρησιμοποιώντας τα συγκεντρωτικά δεδομένα αναζήτησης της Google μπορεί να εκτιμηθεί η ταχύτητα διάδοσης της γρίπης έως και δύο εβδομάδες γρηγορότερα από τα παραδοσιακά συστήματα. Το παράδειγμα αυτό δείχνει πώς το Data Mining μπορεί να μετατρέψει μια μεγάλη συλλογή δεδομένων σε χρήσιμη γνώση που μπορεί να βοηθήσει στην αντιμετώπιση μιας τρέχουσας παγκόσμιας πρόκλησης, **[7], [35]**.

Το Data Mining όπως προαναφέρθηκε μπορεί να θεωρηθεί ως αποτέλεσμα της εξέλιξης της τεχνολογίας των πληροφοριών. Η βιομηχανία βάσεων δεδομένων και διαχείρισης δεδομένων εξελίχθηκε χάρη στην ανάπτυξη πολλών κρίσιμων λειτουργιών όπως η συλλογή δεδομένων και η δημιουργία βάσεων δεδομένων, η διαχείριση δεδομένων (συμπεριλαμβανομένης της αποθήκευσης και ανάκτησης δεδομένων και επεξεργασίας συναλλαγών βάσεων δεδομένων) και η προηγμένη ανάλυση δεδομένων (αποθήκευση δεδομένων και εξόρυξη δεδομένων). Η ανάπτυξη των μηχανισμών συλλογής δεδομένων και δημιουργίας βάσεων δεδομένων αποτέλεσε προϋπόθεση για την μετέπειτα αποτελεσματική ανάπτυξη μηχανισμών για την αποθήκευση και ανάκτηση των δεδομένων καθώς και για την επεξεργασία των ερωτημάτων στις μηχανές αναζήτησης, **[18]**.

Σήμερα, η προηγμένη ανάλυση δεδομένων έχει προχωρήσει στο επόμενο βήμα. Από τη δεκαετία του 1960, οι βάσεις δεδομένων και η τεχνολογία των πληροφοριών εξελίχθηκαν συστηματικά από πρωτόγονα συστήματα επεξεργασίας αρχείων σε εξελιγμένα και ισχυρά συστήματα βάσεων δεδομένων και βάσεων γνώσης. Η έρευνα και η ανάπτυξη συστημάτων βάσεων δεδομένων από τη δεκαετία του 1970 προχώρησε από τα πρώιμα ιεραρχικά συστήματα βάσεων δεδομένων σε συστήματα σχεσιακών βάσεων δεδομένων (όπου τα δεδομένα αποθηκεύονται σε σχεσιακές δομές πίνακα) και σε εργαλεία μοντελοποίησης δεδομένων και μεθόδων πρόσβασης. Επιπλέον, οι χρήστες έχουν αποκτήσει εύκολη και ευέλικτη πρόσβαση σε δεδομένα μέσω των διεπαφών του χρήστη, την βελτιστοποίησης των queries και της διαχείρισης τους .

Οι αποτελεσματικές μέθοδοι επεξεργασίας ηλεκτρονικών συναλλαγών (OLTP), όπου ένα ερώτημα αντιμετωπίζεται ως συναλλαγή μόνο για ανάγνωση, συνέβαλαν ουσιαστικά στην εξέλιξη και την ευρεία αποδοχή της σχεσιακής τεχνολογίας ως σημαντικού εργαλείου για την αποτελεσματική αποθήκευση, ανάκτηση και διαχείριση μεγάλων όγκων δεδομένων από τις βάσεις δεδομένων. Μετά την εγκατάσταση συστημάτων διαχείρισης βάσεων δεδομένων (DBMS), η τεχνολογία βάσεων δεδομένων προχώρησε στην ανάπτυξη προηγμένων συστημάτων βάσεων δεδομένων, αποθήκευσης δεδομένων και εξόρυξης δεδομένων για προηγμένη ανάλυση των δεδομένων και κατανεμημένων βάσεων δεδομένων μέσω διαδικτύου. Τα προηγμένα συστήματα βάσεων δεδομένων, για παράδειγμα, προέκυψαν από την πρόοδο της έρευνας από τα μέσα της δεκαετίας του 1980 και μετά. Αυτά τα συστήματα ενσωματώνουν νέα και ισχυρά μοντέλα δεδομένων όπως τα μοντέλα του εκτεταμένου-σχεσιακού, αντικειμενοστραφούς και αντικείμενο-σχεσιακού

τρόπου οργάνωσης. Τα συστήματα βάσεων δεδομένων με γνώμονα τις εφαρμογές έχουν αναπτυχθεί, συμπεριλαμβάνουν και δεδομένα πολυμέσων, ροών και αισθητήρων, επιστημονικών και μηχανικών βάσεων δεδομένων, βάσεων γνώσεων και βάσεων πληροφοριών για το γραφείο κ.α., **[10], [35]**.

Η προηγμένη ανάλυση δεδομένων ξεκίνησε από τα τέλη της δεκαετίας του 1980. Η εντυπωσιακή πρόοδος της τεχνολογίας του ηλεκτρονικού υπολογιστή στις τελευταίες τρεις δεκαετίες οδήγησε σε μεγάλες προμήθειες ισχυρών και προσιτών υπολογιστών, εξοπλισμό συλλογής δεδομένων και μέσων αποθήκευσης. Αυτή η τεχνολογία προσφέρει μεγάλη ώθηση στη βιομηχανία των βάσεων δεδομένων και της επεξεργασίας πληροφορίας και επιτρέπει τη διαθεσιμότητα μεγάλου αριθμού βάσεων δεδομένων και αποθετηρίων πληροφοριών για τη διαχείριση συναλλαγών, την ανάκτηση πληροφοριών και την ανάλυση δεδομένων.

Τα δεδομένα μπορούν τώρα να αποθηκευτούν σε πολλά διαφορετικά είδη βάσεων δεδομένων και αποθετηρίων πληροφοριών. Μια ανερχόμενη αρχιτεκτονική καταγραφής δεδομένων είναι η αποθήκη δεδομένων (Data Warehouse). Πρόκειται για ένα αποθετήριο πολλαπλών ετερογενών πηγών δεδομένων που οργανώνονται στο πλαίσιο ενοποιημένου σχήματος σε έναν ενιαίο ιστότοπο. Η τεχνολογία αποθήκης δεδομένων περιλαμβάνει τον καθαρισμό δεδομένων, την ενσωμάτωση δεδομένων και την on-line αναλυτική επεξεργασία (OLAP), δηλαδή τεχνικές ανάλυσης με λειτουργίες όπως σύνοψη, ενοποίηση και ενσωμάτωση, καθώς και δυνατότητα προβολής πληροφοριών από διαφορετικές πλευρές. Οι παγκόσμιες βάσεις πληροφοριών που βασίζονται στο Διαδίκτυο, όπως ο Παγκόσμιος Ιστός και διάφορα είδη αλληλένδετων ή ετερογενών βάσεων δεδομένων, έχουν αναπτυχθεί και διαδραματίζουν ζωτικό ρόλο στις μέρες μας. Η αποτελεσματική και αποδοτική ανάλυση των δεδομένων από αυτές τις διαφορετικές μορφές δεδομένων μέσω της ενοποίησης της ανάκτησης πληροφοριών, του Data Mining και των τεχνολογιών ανάλυσης δικτύων πληροφόρησης αποτελεί ένα δύσκολο έργο, **[10], [35]**.

Εικόνα.2.1: Τομείς που προκύπτει το Data Mining

(Πηγή. https://www.sas.com/ru_ua/insights/analytics/data-mining.html)

2.2. Τι είναι το Data Mining

Δεν αποτελεί έκπληξη το γεγονός ότι η εξόρυξη δεδομένων (Data Mining), ως πραγματικά διεπιστημονικό πεδίο, μπορεί να οριστεί με πολλούς διαφορετικούς τρόπους. Ακόμη και ο όρος «εξόρυξη δεδομένων» δεν παρουσιάζει πραγματικά όλα τα κύρια στοιχεία της εικόνας. Αντί του Data Mining θα ήταν περισσότερο κατάλληλη η ονομασία "εξόρυξη γνώσης από δεδομένα" η οποία δυστυχώς είναι κάπως δύσκολη στη χρήση. Ωστόσο ως βραχυπρόθεσμος όρος, η εξόρυξη γνώσης ίσως να μην αντικατοπτρίζει την έμφαση εξόρυξης από μεγάλες ποσότητες δεδομένων. Επιπλέον πολλοί άλλοι όροι έχουν παρόμοια σημασία με την εξόρυξη δεδομένων, όπως για παράδειγμα οι όροι «εξόρυξη γνώσης από δεδομένα», «εξαγωγή γνώσης», «ανάλυση δεδομένων», «αναγνώριση προτύπων», κτλ.

Πολλοί άνθρωποι αντιμετωπίζουν το Data Mining ως συνώνυμο ενός άλλου ευρέως χρησιμοποιούμενου όρου, της ανακάλυψης γνώσης από τα δεδομένα, ενώ άλλοι θεωρούν το Data Mining ως απλό αλλά σημαντικό βήμα στη διαδικασία της ανακάλυψης της γνώσης. Η διαδικασία του Data Mining είναι και μπορεί να θεωρηθεί ως επαναληπτική ακολουθία των ακόλουθων βημάτων:

- ✓ Καθαρισμός δεδομένων (για την εξάλειψη των ασυνεπών εντός των δεδομένων)
- ✓ Ενσωμάτωση δεδομένων (όπου μπορούν να συνδυαστούν πολλαπλές πηγές δεδομένων)

✓ Επιλογή δεδομένων (όπου δεδομένα που σχετίζονται με το στόχο της δεδομένης ανάλυσης ανακτώνται από το βάση δεδομένων)

✓ Μετασχηματισμός δεδομένων (όπου τα δεδομένα μετατρέπονται και ενοποιούνται σε μορφή κατάλληλη για εξόρυξη διενεργώντας πράξεις περίληψης ή συγκέντρωσης / συμπίεσης / συμπύκνωσης)

✓ Εξόρυξη δεδομένων (μια βασική διαδικασία όπου εφαρμόζονται έξυπνες μέθοδοι για την εξαγωγή πρότυπων δεδομένων)

✓ Αξιολόγηση μοτίβων (για να προσδιοριστούν τα πραγματικά ενδιαφέροντα πρότυπα που αντιπροσωπεύουν τη γνώση με βάση το εκάστοτε ενδιαφέρον του αναλυτή)

✓ Παρουσίαση της γνώσης (όπου οι τεχνικές απεικόνισης και εκπροσώπησης της γνώσης χρησιμοποιούνται για την παρουσίαση εξειδικευμένων γνώσεων στους χρήστες), **[16], [35].**

Τα πρώτα βήματα είναι διαφορετικές μορφές προεπεξεργασίας δεδομένων, που προετοιμάζουν τα δεδομένα για εξόρυξη. Το βήμα εξόρυξης δεδομένων μπορεί να αλληλεπιδρά με τον χρήστη ή με μια βάση γνώσεων. Τα δεδομένα παρουσιάζονται στον χρήστη και μπορούν να αποθηκευτούν ως νέες γνώσεις στην παλιά βάση γνώσεων. Η προηγούμενη υποενότητα αναδεικνύει την εξόρυξη δεδομένων ως ένα βήμα στη διαδικασία της ανακάλυψης γνώσης (knowledge discovery, KD), **[7], [10]**.

Ωστόσο, αν και στη βιομηχανία και στο περιβάλλον της Έρευνας χρησιμοποιείται συχνά ο όρος Data Mining, θα πρέπει να ανατρέξει κανείς σε ολόκληρη τη διαδικασία της ανακάλυψης της γνώσης (ίσως επειδή ο όρος είναι υποσύνολο της ανακάλυψης γνώσεων από τα δεδομένα). Επομένως, υιοθετείται εδώ μια ευρεία θεώρηση του Data Mining: Το Data Mining είναι η διαδικασία της ανεύρεσης ενδιαφερόντων προτύπων και γνώσης από μεγάλους όγκους δεδομένων. Οι πηγές δεδομένων μπορούν να περιλαμβάνουν βάσεις δεδομένων, δεδομένα από μεγάλες αποθήκες δεδομένων (Data Warehouse), το Διαδίκτυο, άλλα αποθετήρια πληροφοριών ή δεδομένα που μεταδίδονται στο δίκτυο ενός συστήματος δυναμικά.

2.3. Δεδομένα που αξιοποιούνται στο Data Mining

2.3.1. Βάσεις Δεδομένων

Ως γενική τεχνολογία, το Data Mining μπορεί να εφαρμοστεί σε οποιοδήποτε είδος δεδομένων, εφόσον τα δεδομένα έχουν νόημα για μια εφαρμογή-στόχο (Εικόνα 2.2). Οι πιο βασικές μορφές δεδομένων πρόσφορων για Data Mining είναι (α) από βάσεις δεδομένων, (β) από αποθήκες δεδομένων, και (γ) από ανταλλαγές δεδομένων. Οι έννοιες και οι τεχνικές που παρουσιάζονται εδώ, εστιάζουν σε αυτά τα δεδομένα. Το Data Mining μπορεί επίσης να εφαρμοστεί σε άλλες μορφές δεδομένων, όπως ροές δεδομένων, δεδομένα παραγγελιών / ακολουθιών, δεδομένα γραφικών ή δικτυωμένων δεδομένων, χωρικά δεδομένα, δεδομένα κειμένου και δεδομένα πολυμέσων. Το Data Mining σίγουρα θα συνεχίσει να αγκαλιάζει νέους τύπους δεδομένων καθώς αυτοί θα εμφανίζονται.

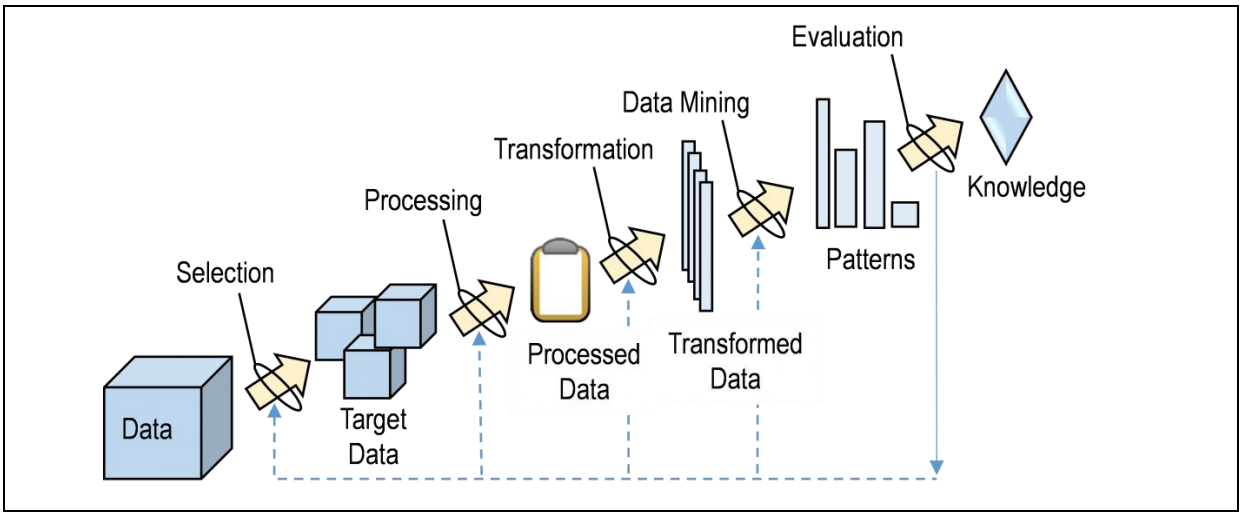
Ένα **σύστημα βάσης δεδομένων**, που ονομάζεται επίσης σύστημα διαχείρισης βάσεων δεδομένων (DBMS), αποτελείται από μια συλλογή αλληλένδετων δεδομένων, γνωστή ως βάση δεδομένων, και ένα σύνολο προγραμμάτων λογισμικού για την πρόσβαση στα δεδομένα και τη διαχείρισή τους. Τα προγράμματα λογισμικού παρέχουν μηχανισμούς το σχεδιασμό και την εισαγωγή στοιχείων στις βάσεις δεδομένων στις αποθήκες δεδομένων. Για τον προσδιορισμό και τη διαχείριση ταυτόχρονων προσβάσεων καθώς και για τη διασφάλιση της συνέπειας και της ασφάλειας των πληροφοριών, τα δεδομένα αποθηκεύονται παρά τις συγκρούσεις ή τις προσπάθειες μη εξουσιοδοτημένης πρόσβασης, **[10], [15]**.

Μια σχεσιακή βάση δεδομένων (relational database) είναι μια συλλογή από πίνακες, σε κάθε έναν από τους οποίους έχει εκχωρηθεί ένα μοναδικό όνομα. Κάθε πίνακας αποτελείται από ένα σύνολο χαρακτηριστικών (στήλες ή πεδία) και συνήθως σ' αυτόν αποθηκεύεται ένα μεγάλο σύνολο πλειάδων (εγγραφές ή σειρές). Κάθε πλειάδα σε ένα σχεσιακό πίνακα αντιπροσωπεύει ένα αντικείμενο (π.χ. άτομο, προϊόν, κλπ.) που προσδιορίζεται από ένα μοναδικό κλειδί και περιγράφεται από ένα σύνολο τιμών των χαρακτηριστικών του. Για τις σχεσιακές βάσεις δεδομένων συνήθως αρχικά κατασκευάζεται ένα μοντέλο δεδομένων, όπως π.χ. ένα μοντέλο δεδομένων οντότητας-σχέσης ή οντοτήτων-συσχετίσεων (ER).. Ένα μοντέλο δεδομένων ER αντιπροσωπεύει τη

βάση δεδομένων ως το σύνολο των οντοτήτων που θα αποθηκευτούν σ' αυτήν μαζί με τις σχέσεις μεταξύ τους.

Τα σχεσιακά δεδομένα μπορούν να προσεγγιστούν από ερωτήματα queries που ο χρήστης απευθύνει προς τη βάση, διατυπωμένα σε μια σχεσιακή γλώσσα (π.χ. γλώσσα SQL – Structured Query Language) ή με τη βοήθεια γραφικών διεπαφών, (GUI – Graphical User Interfaces). Ένα query είναι διατυπωμένο αρχικά σε μία γλώσσα σχεδόν κατανοητή (και) από τον άνθρωπο, αλλά μετασχηματίζεται σε μια σειρά από σχεσιακές λειτουργίες που θα εκτελεστούν μέσα στη βάση δεδομένων και θα εφαρμοστούν πάνω στα δεδομένα της, όπως σύνδεση, επιλογή και προβολή, και στη συνέχεια βελτιστοποίηση για αποτελεσματική επεξεργασία. Ένα query επιτρέπει την ανάκτηση συγκεκριμένων υποσυνόλων των δεδομένων. Αν υποτεθεί ότι ο στόχος του χρήστη είναι να αναλύσει τα δεδομένα της βάσης για να εξάγει την απάντηση σε ένα ερώτημα διατυπωμένο σε υψηλό επίπεδο (με σχετικά αφηρημένους όρους). Μέσα από τη χρήση των ερωτημάτων (queries), μπορεί να ζητήσει πράγματα όπως: «Δείξε μου μια λίστα με όλα τα στοιχεία που πωλούνται κατά το τελευταίο εξάμηνο» κτλ. Οι ερωτήσεις αυτές μπορούν να ενεργοποιήσουν και άλλες λειτουργίες όπως το άθροισμα, ο υπολογισμός μέσης τιμής, μέγιστου και ελάχιστου, κτλ.

Όταν από τις σχεσιακές βάσεις δεδομένων εξαχθούν τα δεδομένα που αντιστοιχούν στο συγκεκριμένο πρόβλημα ή ερώτημα-στόχο, ο αναλυτής μπορεί να προχωρήσει περαιτέρω αναζητώντας τάσεις (trends) ή μοντέλα (models) εντός των εξαχθέντων δεδομένων. Για παράδειγμα, τα συστήματα Data Mining μπορούν να αναλύσουν τα δεδομένα των πελατών π.χ. μιας τράπεζας για να προβλέψουν τον πιστωτικό κίνδυνο νέων πελατών με βάση το εισόδημα, την ηλικία και την προηγούμενη πίστωση κτλ. Τα συστήματα Data Mining ενδέχεται επίσης να ανιχνεύουν ακραίες αποκλίσεις από τις τυπικές συμπεριφορές, δηλαδή στοιχεία με πωλήσεις που απέχουν πολύ από τις αναμενόμενες σε σχέση με το προηγούμενο έτος τιμές. Για παράδειγμα το Data Mining μπορεί να ανακαλύψει ότι υπάρχει μια αλλαγή στη συσκευασία ενός αντικειμένου ή μια σημαντική αύξηση της τιμής. Οι σχεσιακές βάσεις δεδομένων είναι μία από τις πιο διαδεδομένες και πλουσιότερες μορφές αποθετηρίων και έτσι είναι και μια σημαντική πηγή δεδομένων στην υπηρεσία του Data Mining, **[10], [17]**.



Εικόνα.2.2: Διαδρομή Data Mining από την Βάση Δεδομένων έως την γνώση

(Πηγή. https://behavior.lbl.gov/?q=node/11)

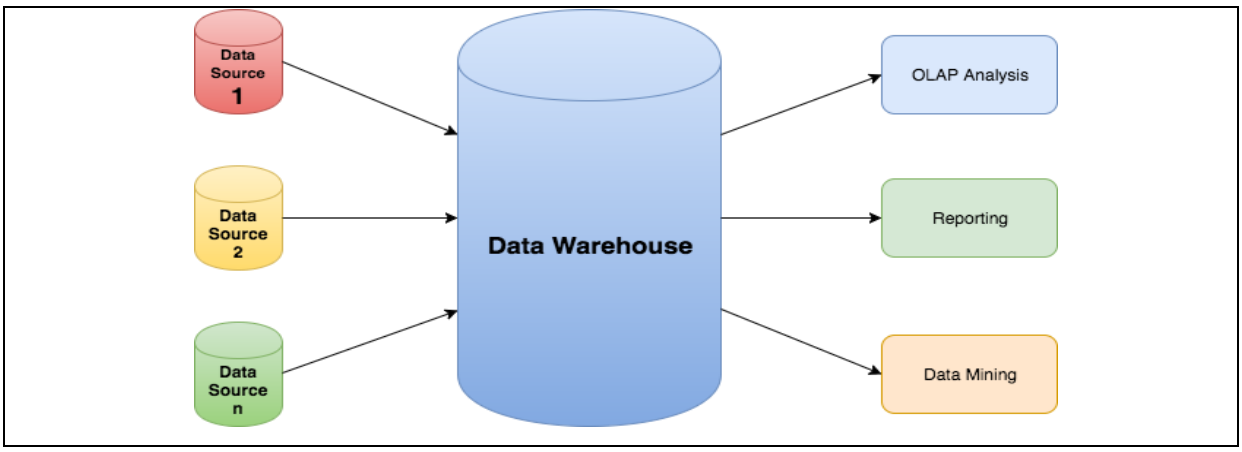
2.3.2. Αποθήκες ή Αποθετήρια Δεδομένων (Data Warehouses)

Ας υποτεθεί μια επιτυχημένη διεθνής εταιρεία με υποκαταστήματα σε όλο τον κόσμο. Κάθε υποκατάστημα έχει το δικό του σύνολο βάσεων δεδομένων. Ο πρόεδρος της εταιρείας ζητά να αποδοθεί μια ανάλυση των πωλήσεων της εταιρείας ανά είδος και ανά υποκατάστημα για το τρίτο τρίμηνο του έτους. Πρόκειται για ένα δύσκολο έργο, ιδίως επειδή τα σχετικά στοιχεία βρίσκονται σε αρκετές και διαφορετικές βάσεις δεδομένων που βρίσκονται φυσικά κατανεμημένες σε πολυάριθμες τοποθεσίες. Εάν η εταιρεία είχε μια αποθήκη δεδομένων (Data Warehouse), αυτή η εργασία θα ήταν εύκολη. Μια αποθήκη δεδομένων είναι ένα αποθετήριο πληροφοριών που συλλέγονται από πολλαπλές πηγές, αποθηκεύονται σε ένα ενοποιημένο σύστημα και συνήθως παραμένουν σε έναν ενιαίο ιστότοπο. Οι αποθήκες δεδομένων κατασκευάζονται με την διαδικασία επεξεργασίας δεδομένων, ολοκλήρωσης δεδομένων, μετασχηματισμού δεδομένων, φόρτωσης δεδομένων και περιοδικής ανανέωσης δεδομένων, **[7], [15]**.

Για να διευκολυνθεί η λήψη αποφάσεων, τα δεδομένα σε μια αποθήκη δεδομένων (Data Warehouse) είναι οργανωμένα (π.χ. πελάτης, στοιχείο, προμηθευτής και δραστηριότητα). Τα δεδομένα αποθηκεύονται για να παρέχουν πληροφορίες από ιστορική άποψη, όπως π.χ. κατά τους τελευταίους 6 έως 12 μήνες, και συνήθως συνοψίζουν τα γενικά στοιχεία. Για παράδειγμα, αντί να αποθηκευθούν οι λεπτομέρειες κάθε συναλλαγής

πώλησης αποθηκεύεται μια σύνοψη των συναλλαγών ανά τύπο πωλούμενου προϊόντος για κάθε περιοχή πωλήσεων και για δεδομένη χρονική περίοδο.

Μια αποθήκη δεδομένων συνήθως στηρίζεται σε μια πολυδιάστατη δομή δεδομένων, που ονομάζεται «κύβος δεδομένων», στον οποίο κάθε διάσταση αντιστοιχεί σε ένα χαρακτηριστικό ή σε ένα σύνολο χαρακτηριστικών στο σχήμα, και κάθε κύτταρο αποθηκεύει την τιμή μέτρου κάποιου συνολικού μεγέθους, όπως ο πολυδιάστατος χώρος μέτρησης σε ένα σύστημα OLAP. Συνεπώς, επιτρέπει την εξερεύνηση πολλαπλών συνδυασμών διαστάσεων σε ποικίλα επίπεδα στο Data Mining, και έτσι έχει μεγαλύτερες δυνατότητες να ανακαλύψει ενδιαφέροντα πρότυπα που αντιπροσωπεύουν τη γνώση. Μια βασική δομή ενός Data Warehouse φαίνεται στη επόμενη εικόνα(Εικόνα 2.3).



Εικόνα.2.3: Δομή του Data Warehouse

(Πηγή. http://www.lastnightstudy.com/Show?id=31/Data-Warehouse)

2.3.3. Άλλα Δεδομένα

Εκτός από τα δεδομένα μίας σχεσιακής βάσης δεδομένων (Relational Data Base) και τα δεδομένα μίας αποθήκης δεδομένων (Data Warehouse), υπάρχουν πολλά άλλα είδη δεδομένων που έχουν ευπροσάρμοστες μορφές και δομές και μάλλον διαφορετικές σημασιολογικά έννοιες. Αυτά τα είδη δεδομένων μπορούν να παρατηρηθούν σε πολλές εφαρμογές: σχετίζονται με το χρόνο, όπως π.χ. τα δεδομένα αλληλουχίας (ιστορικά αρχεία, δεδομένα Χρηματιστηρίου και οικονομικές χρονοσειρές, αλληλουχίες βιολογικών δεδομένων), ροές δεδομένων (π.χ. δεδομένα παρακολούθησης βίντεο και εν γένει μετρήσεων από αισθητήρα, τα οποία είναι συνεχώς μεταδιδόμενα – streaming), χωρικά

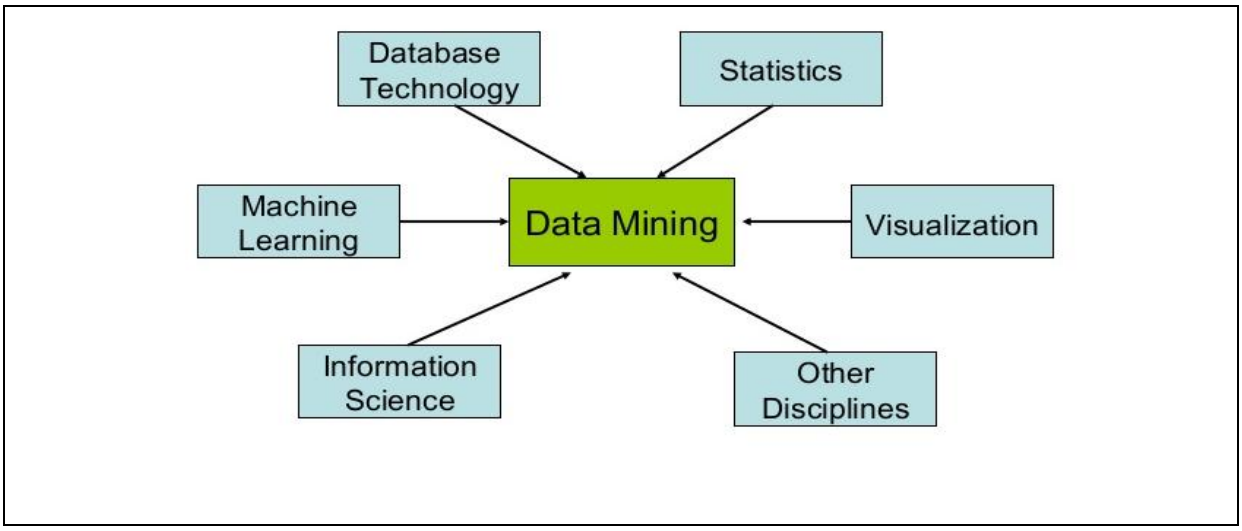
δεδομένα (π.χ. χάρτες), δεδομένα μηχανικής σχεδίασης (π.χ. σχεδιασμός κτιρίων, εξαρτημάτων του συστήματος ή ολοκληρωμένων κυκλωμάτων), δεδομένα υπερκειμένου (hypertext) και δεδομένα πολυμέσων (multimedia), συμπεριλαμβανομένων των δεδομένων κειμένου, εικόνων, βίντεο και ήχου, γραφήματα και δεδομένα δικτύου (π.χ. κοινωνικά και πληροφοριακά δίκτυα) και παγκόσμιου ιστού (μια τεράστια, ευρέως κατανεμημένη αποθήκη πληροφοριών που διατίθεται μέσω του Διαδικτύου). Αυτές οι εφαρμογές φέρνουν νέες προκλήσεις, όπως ο τρόπος χειρισμού δεδομένων που μεταφέρουν ειδικές δομές (π.χ. αλληλουχίες, δέντρα, γραφήματα και δίκτυα) και συγκεκριμένη σημασιολογία (όπως παραγγελία, εικόνα, ήχος και βίντεο, περιεχόμενο και συνδεσιμότητα), **[7], [10]**.

Διάφορα είδη γνώσεων μπορούν να εξορύσσονται από αυτά τα είδη δεδομένων. Εδώ θα αναφερθούν μόνο μερικά. Όσον αφορά τα χρονικά δεδομένα, για παράδειγμα, μπορούν να οριστούν τραπεζικά δεδομένα που εντοπίζουν μεταβολές («τάσεις» - trends) και που ενδέχεται να βοηθήσουν στον προγραμματισμό των τραπεζικών ταμείων, σύμφωνα με τον όγκο των συναλλαγών επισκεψιμότητα πελατών, κτλ. Τα στοιχεία της χρηματιστηριακής αγοράς μπορούν να εξορύσσονται για να αποκαλύψουν τάσεις που θα μπορούσαν να βοηθήσουν επενδυτικές στρατηγικές (π.χ. τον καλύτερο χρόνο για να αγοραστεί το απόθεμα μιας εταιρείας). Με τα χωρικά δεδομένα, μπορούν να αναζητηθούν μοτίβα που περιγράφουν αλλαγές π.χ. στα ποσοστά φτώχειας που βασίζονται στις αποστάσεις των πόλεων από τους μεγάλους αυτοκινητοδρόμους. Μπορεί επίσης να εξεταστεί ένα σύνολο από χωρικά αντικείμενα προκειμένου να ανακαλυφθεί ποια υποσύνολα αντικειμένων είναι χωρικά αυτοσυσχετισμένα ή συσχετισμένα μεταξύ τους. Με την εξόρυξη δεδομένων κειμένου (Text mining), όπως η δείχνει η έντονα αυξανόμενη βιβλιογραφία σχετικά με τα δεδομένα αυτά κατά τα τελευταία δέκα χρόνια, μπορούν να εντοπιστούν τα επίκαιρα θέματα ανταλλαγών και συζητήσεων μεταξύ των χρηστών των κοινωνικών, π.χ., δικτύων σε δεδομένη στιγμή ή χρονική περίοδο. Σε εκπαιδευτικό πλαίσιο, με βάση τα σχόλια των χρηστών σχετικά με τα μαθήματα (τα οποία συχνά υποβάλλονται ως σύντομα μηνύματα κειμένου) μπορούν να αξιολογηθούν τα συναισθήματα των φοιτητών και να προκύψει πόσο καλά ανταποκρίνεται εκπαιδευτικά ένα μάθημα στους φοιτητές. Από τα δεδομένα πολυμέσων, μπορούν να ληφθούν εικόνες για να εντοπιστούν αντικείμενα και να ταξινομηθούν σε κατηγορίας, με την ανάθεση σημασιολογικών ετικετών στο καθένα. Με την εξόρυξη δεδομένων βίντεο ενός ψηφιακού παιχνιδιού, μπορούν να ανιχνευθούν ακολουθίες βίντεο που αντιστοιχούν σε

συγκεκριμένους στόχους. Η εξόρυξη δεδομένων παγκόσμιου ιστού μπορεί να βοηθήσει να αποτυπωθεί η διανομή πληροφοριών στο διαδίκτυο γενικότερα, καθώς και να χαρακτηριστούν και να ταξινομηθούν οι ιστοσελίδες, αποκαλύπτοντας έτσι τη δυναμική του παγκόσμιου ιστού μέσω της συσχέτισής τους.

2.4. Τεχνολογίες που χρησιμοποιούνται στο Data Mining

Ως ένα νέο πεδίο με έντονα εφαρμοσμένο – πρακτικό χαρακτήρα, το Data Mining έχει ενσωματώσει πολλές τεχνικές από άλλες επιστήμες ή επιστημονικές περιοχές, όπως η στατιστική, η μηχανική μάθηση, η αναγνώριση προτύπων, οι βάσεις δεδομένων και τα συστήματα αποθήκης δεδομένων, η ανάκτηση πληροφοριών, η ψηφιοποίηση και οπτικοποίηση δεδομένων και οι αλγόριθμοι υψηλής απόδοσης, μεταξύ άλλων(Εικόνα 2.4). Η διεπιστημονική φύση της έρευνας και της ανάπτυξης στον τομέα του Data Mining συμβάλλει σημαντικά στην επιτυχία του και στη διαρκή επέκταση των πεδίων εφαρμογής του. Στη συνέχεια θα δοθούν ορισμένα βασικά παραδείγματα των διαφόρων κλάδων που επηρεάζουν έντονα την ανάπτυξη μεθόδων εξόρυξης δεδομένων με συνοπτικό τρόπο.



Εικόνα.2.4: Τεχνολογίες Data Mining

(Πηγή. https://www.slideshare.net/sanjaypaularvind/lecture-data-mining)

2.4.1. Στατιστική

Η Στατιστική μελετά τη συλλογή, ανάλυση, ερμηνεία ή εξήγηση και παρουσίαση των δεδομένων. Το Data Mining έχει άμεση σχέση με τα στατιστικά στοιχεία. Ένα στατιστικό μοντέλο είναι ένα σύνολο μαθηματικών λειτουργιών που περιγράφουν τη συμπεριφορά των αντικειμένων, οντοτήτων ή συστημάτων που εξετάζονται, από την άποψη των τυχαίων μεταβλητών (stochastic processes) και της σχετικής κατανομής πιθανότητάς τους (probability distribution). Τα στατιστικά μοντέλα χρησιμοποιούνται ευρέως για τον υπολογισμό των κατηγοριών των δεδομένων (classification), **[7], [10], [17]**.

Για παράδειγμα, σε εργασίες Data Mining όπως ο χαρακτηρισμός και η ταξινόμηση δεδομένων, μπορούν να δημιουργηθούν στατιστικά μοντέλα κατηγοριών και στόχων. ως αποτέλεσμα μιας εργασίας Data Mining. Εναλλακτικά, οι εργασίες εξόρυξης δεδομένων μπορούν να χτιστούν πάνω στα στατιστικά μοντέλα. Για παράδειγμα, μπορούν να χρησιμοποιηθούν στατιστικά στοιχεία για να διαμορφωθούν/συμπληρωθούν με τον καλύτερο/φυσικότερο τρόπο ορισμένα δεδομένα που λείπουν. Στη συνέχεια, το Data Mining μπορεί να χρησιμοποιήσει το μοντέλο για τον εντοπισμό και τη διαχείριση των τιμών στα δεδομένα που λείπουν.

Η στατιστική έρευνα αναπτύσσει εργαλεία πρόβλεψης (prediction) με τη χρήση δεδομένων και στατιστικών μοντέλων. Οι στατιστικές μέθοδοι μπορούν να χρησιμοποιηθούν για να συνοψίσουν ή να περιγράψουν μια συλλογή δεδομένων. Οι στατιστική είναι χρήσιμη για την εξόρυξη διαφόρων μορφών δεδομένων όπως για την κατανόηση του μηχανισμού που δημιουργεί και επηρεάζει τα μοτίβα (patterns). Συμπερασματικά στατιστικά στοιχεία (ή πρόβλεψη στατιστικών στοιχείων) υποδεικνύουν δεδομένα με τρόπο που να αποδίδουν τυχαία και αβέβαιη εικόνα στις παρατηρήσεις, και χρησιμοποιούνται για να συνάγουν συμπεράσματα σχετικά με τη διαδικασία μιας έρευνας.

Οι στατιστικές μέθοδοι μπορούν επίσης να χρησιμοποιηθούν για την επαλήθευση των αποτελεσμάτων εξόρυξης δεδομένων. Για παράδειγμα, αφού μία διαδικασία Data Mining καταλήξει σε ένα μοντέλο ταξινόμησης ή πρόβλεψης, το μοντέλο πρέπει να επαληθευθεί με στατιστικά στοιχεία που να έχουν προκύψει από μετρήσεις πάνω σε πραγματικές/φυσικές διαδικασίες. Ένας στατιστικός έλεγχος υποθέσεων (hypothesis testing) (μερικές φορές αποκαλείται επιβεβαίωση της ανάλυσης δεδομένων – data validation) λαμβάνει στατιστικές αποφάσεις χρησιμοποιώντας πειραματικά δεδομένα. Ένα

αποτέλεσμα ονομάζεται στατιστικά σημαντικό εάν είναι πολύ πιθανό να συμβεί. Αν η ταξινόμηση ή πρόβλεψη του μοντέλου ισχύει, τότε οι περιγραφικές στατιστικές του μοντέλου αυξάνουν την αξιοπιστία του μοντέλου. Η εφαρμογή στατιστικών μεθόδων στην εξόρυξη δεδομένων είναι μεγάλη. Συχνά μια σοβαρή πρόκληση για τους ερευνητές είναι πώς να κλιμακώσουν μια στατιστική μέθοδο ώστε να λειτουργεί με ένα ιδιαίτερα μεγάλο σύνολο δεδομένων (big data), **[7], [10], [17]**.

Οι στατιστικές μέθοδοι έχουν μεγάλη υπολογιστική πολυπλοκότητα (computational complexity). Όταν εφαρμόζονται τέτοιες μέθοδοι σε ιδιαίτερα μεγάλα σύνολα δεδομένων, ενδεχομένως κατανεμημένων σε πολλαπλές λογικές ή φυσικές τοποθεσίες, θα πρέπει να σχεδιάζονται προσεκτικά και να ρυθμιστούν ώστε να μειωθεί το υπολογιστικό κόστος. Αυτή η πρόκληση καθίσταται ακόμα πιο έντονη για εφαρμογές λογισμικού που «τρέχουν» πάνω από το Διαδίκτυο, όπως οι προτάσεις ηλεκτρονικών ερωτήσεων στο μηχανές αναζήτησης, όπου η εξόρυξη δεδομένων είναι απαραίτητη ώστε η εφαρμογή να χειρίζεται διαρκώς και γρήγορα, σε πραγματικό χρόνο, τις ροές δεδομένων.

2.4.1.1. Προσεγγιστικές Στατιστικές Μέθοδοι

Εκτός από τις στατιστικές μεθόδους που οδηγούν σε συγκέντρωση, συμπίεση ή συμπύκνωση και περιληπτική αναπαράσταση των δεδομένων, υπάρχουν και οι στατιστικές μέθοδοι για την ανίχνευση των εξωστρεφών παραδοχών σχετικά με την ομαλότητα των δεδομένων. Υποτίθεται ότι τα κανονικά αντικείμενα μέσα σε ένα σύνολο δεδομένων δημιουργούνται από μια στοχαστική διαδικασία (π.χ. ένα γενετικό μοντέλο). Κατά συνέπεια, κανονικά αντικείμενα εμφανίζονται σε περιοχές όπου το στοχαστικό μοντέλο δίνει μεγάλη πιθανότητα ενώ υπερβολικά αντικείμενα εμφανίζονται στις περιοχές χαμηλής πιθανότητας.

Η γενική ιδέα πίσω από τις στατιστικές μεθόδους ανίχνευσης των εξωστρεφών παραδοχών είναι να εκπαιδευθεί ένα γενετικό μοντέλο με βάση σειρά από διαθέσιμα παραδείγματα δεδομένων, προσαρμόζοντας το σύνολο δεδομένων κατάλληλα και στη συνέχεια προσδιορίζοντας τα αντικείμενα με χαμηλή πιθανότητα σε περιοχές του μοντέλου ως υπερβολικές τιμές. Ωστόσο, υπάρχουν πολλοί διαφορετικοί τρόποι και αλγόριθμοι για να εκπαιδευθούν τα γενετικά μοντέλα. Γενικά οι στατιστικές μέθοδοι ανίχνευσης των εξωστρεφών παραδοχών μπορούν να χωριστούν σε δύο βασικές κατηγορίες: τις

παραμετρικές μεθόδους και τις μη παραμετρικές μεθόδους, σύμφωνα με τον τρόπο με τον οποίο τα μοντέλα καθορίζονται.

2.4.2. Μηχανική Μάθηση (Machine Learning)

Η μηχανική μάθηση (machine learning) διερευνά τον τρόπο με τον οποίο οι υπολογιστές μπορούν να "μάθουν" (να εκπαιδευθούν ώστε να εκτελούν επιτυχώς συγκεκριμένη εργασία-στόχο), με βάση τα δεδομένα. Ένας κύριος τομέας έρευνας αφορά την ανάπτυξη λογισμικών ηλεκτρονικών υπολογιστών που εκπαιδεύεται ώστε αυτόματα να αναγνωρίζει περίπλοκα πρότυπα και να λαμβάνουν έξυπνες αποφάσεις με βάση τα δεδομένα εκπαίδευσής τους. Για παράδειγμα, ένα τυπικό πρόβλημα μηχανικής μάθησης είναι να προγραμματιστεί ένας υπολογιστής ώστε αυτόματα να αναγνωρίζει τους χειρόγραφους ταχυδρομικούς κώδικες στο ταχυδρομείο. Η μηχανική μάθηση είναι ένας ταχέως αναπτυσσόμενος τομέας. Εδώ, παρουσιάζονται κλασικά προβλήματα μηχανικής μάθησης, που σχετίζονται ιδιαίτερα με την εξόρυξη δεδομένων.

✓ **Η εποπτευόμενη μάθηση** (supervised learning) είναι βασικά ένα συνώνυμο της ταξινόμησης (classification). Η εποπτεία στη μάθηση σημαίνει ότι κατά τη φάση της εκπαίδευσης της «μηχανής» (του λογισμικού), εισάγονται σ' αυτό επισημασμένα παραδείγματα που αποτελούν το σύνολο των δεδομένων εκπαίδευσης (training set). Για παράδειγμα, στο πρόβλημα αναγνώρισης ταχυδρομικού κώδικα, θα μπορούσε να δημιουργηθεί ένα σύνολο εικόνων χειρόγραφου ταχυδρομικού κώδικα, μέσω scanner, καθεμία εικόνα συνοδευόμενη από την αντίστοιχη μηχανικά αναγνώσιμη (ορθή) «μετάφρασή» της. Τα ζεύγη χειρόγραφων εικόνων-ορθών αναγνώσεών τους, στη συνέχεια χρησιμοποιούνται ως παραδείγματα εκπαίδευσης, που εισάγονται επαναληπτικά στον αλγόριθμο εκπαίδευσης, έως ότου ελαχιστοποιηθεί το σφάλμα ανάμεσα στην ορθή και στην εκάστοτε προκύπτουσα από τον αλγόριθμο ανάγνωση κάθε χειρόγραφης εικόνας. Τότε θεωρείται ότι ολοκληρώθηκε επιτυχώς η «φάση εκμάθησης» του μοντέλου ταξινόμησης. Στη συνέχεια, στην κυρίως φάση λειτουργίας του αλγόριθμου αυτόματης κατηγοριοποίησης, και στο βαθμό που η φάση εκπαίδευσης υπήρξε επιτυχημένη, αντιμετωπίζει άγνωστες εικόνες εισόδου και τις κατηγοριοποιεί ορθά με μεγάλο ποσοστό επιτυχίας (πιθανότητα ορθής κατηγοριοποίησης), **[7], [35]**.

✓ **Η μη επιτηρούμενη μάθηση** (unsupervised learning) είναι ουσιαστικά συνώνυμο της ομαδοποίησης (clustering). Η διαδικασία μάθησης δεν εποπτεύεται, αφού τα παραδείγματα εισόδου δεν φέρουν επισήμανση της ορθής κλάσης του καθενός. Ο αλγόριθμος τα ομαδοποιεί μόνος του με βάση κάποιο κριτήριο ομοιότητας / ανομοιότητας («απόστασης»). Συνήθως, ομαδοποίηση χρησιμοποιείται για να ανακαλυφθούν ομάδες (clusters) μέσα στα δεδομένα. Για παράδειγμα, σε μια περίπτωση μηχανικής μάθησης χωρίς επίβλεψη, αυτή η μέθοδος μπορεί να λάβει ως είσοδο ένα σύνολο εικόνων χειρόγραφων αριθμητικών ψηφίων από το 0 έως το 9. Ας υποτεθεί ότι η μέθοδος ανακαλύπτει ότι ο καλύτερος τρόπος να ομαδοποιηθούν τα δεδομένα (εικόνες) είναι σε δέκα (10) ομάδες. Αυτά τα δέκα clusters μπορεί πράγματι να αντιστοιχούν στα 10 διαφορετικά ψηφία από 0 έως 9, αντίστοιχα, αν ο αλγόριθμος έχει λειτουργήσει σωστά. Ωστόσο, δεδομένου ότι τα δεδομένα εκπαίδευσης δεν φέρουν ετικέτα της ορθής κατηγορίας, το μαθησιακό αυτό μοντέλο, ακόμα και όταν λειτουργεί επιτυχημένα, δεν μπορεί να δώσει (δηλώσει) τη σημασία των ομάδων που ανακάλυψε, **[7], [35]**.

✓ **Η ημι-εποπτευόμενη μάθηση** είναι μια κατηγορία τεχνικών μηχανικής μάθησης που κάνουν χρήση τόσο της επισήμανσης των παραδειγμάτων εισόδου με τις ορθές ετικέτες όσο και των μη επισημασμένων παραδειγμάτων κατά την εκπαίδευση ενός μοντέλου. Τα επισημασμένα παραδείγματα χρησιμοποιούνται για να αντιληφθεί το μοντέλο το πλήθος των τάξεων ενώ εντός των τάξεων χρησιμοποιούνται παραδείγματα χωρίς επισημάνσεις. Για ένα πρόβλημα δύο κατηγοριών, μπορεί να το σύνολο παραδειγμάτων που ανήκουν σε μία κατηγορία ως «θετικά» παραδείγματα και αυτά που ανήκουν στην άλλη κατηγορία ως «αρνητικά» παραδείγματα. Χρησιμοποιώντας τα μη επισημασμένα παραδείγματα, μπορεί να τελειοποιηθεί το «σύνορο» μεταξύ θετικής και αρνητικής απόφασης του σε μια σταθερή γραμμή, **[7], [35]**.

✓ **Η ενεργός μάθηση** (active learning) είναι μια προσέγγιση μηχανής μάθησης που επιτρέπει στους χρήστες να διαδραματίσουν ενεργό ρόλο στη διαδικασία εκπαίδευσης του αλγορίθμου. Μια προσέγγιση ενεργής μάθησης μπορεί να ζητήσει από έναν χρήστη μια πληροφορία, η οποία μπορεί να προέρχεται είτε από ένα σύνολο μη επισημασμένων παραδειγμάτων είτε από το πρόγραμμα εκμάθησης. Ο στόχος είναι να βελτιστοποιηθεί η ποιότητα του μοντέλου ώστε να αποκτήσει «γνώση» από τους ανθρώπους-χρήστες αλληλεπιδρώντας μαζί τους, **[7], [35]**.

2.4.3. Data Bases και Data Warehouses

Η έρευνα των συστημάτων βάσεων δεδομένων επικεντρώνεται στη δημιουργία, συντήρηση και χρήση των βάσεων δεδομένων για οργανισμούς και τελικούς χρήστες. Συγκεκριμένα, η έρευνα στο πεδίο αυτό έχει αναπτύξει μηχανισμούς σχεδίασης συστημάτων βάσεων δεδομένων, αναγνωρισμένες αρχές και μοντέλα δεδομένων, γλώσσες διατύπωσης ερωτήσεων (query languages), μεθόδους επεξεργασίας των ερωτημάτων και μεθόδους βελτιστοποίησης για την εξαγωγή των απαντήσεων, την αποθήκευση δεδομένων και τις μεθόδους ευρετηρίασης και πρόσβασης στα δεδομένα της βάσης. Τα συστήματα βάσεων δεδομένων είναι γνωστά για την υψηλή επεκτασιμότητα τους στην επεξεργασία πολύ μεγάλων, σχετικά δομημένων συνόλων δεδομένων.

Πολλές εργασίες εξόρυξης δεδομένων πρέπει να χειρίζονται μεγάλα σύνολα δεδομένων ή / και σε πραγματικό χρόνο, με γρήγορη ροή δεδομένων. Επομένως η εξόρυξη δεδομένων μπορεί να αξιοποιήσει τις τεχνολογίες κλιμακούμενων βάσεων δεδομένων για την επίτευξη υψηλής απόδοσης και την επεκτασιμότητα σε μεγάλα σύνολα δεδομένων. Επιπλέον, οι εργασίες Data Mining μπορούν να χρησιμοποιηθούν για την επέκταση της ικανότητας των υφιστάμενων συστημάτων βάσεων δεδομένων να ικανοποιήσουν προηγμένες ανάγκες των χρηστών και εξελιγμένες απαιτήσεις ανάλυσης δεδομένων.

Πρόσφατα συστήματα βάσεων δεδομένων έχουν αποκτήσει δυνατότητες για συστηματική ανάλυση των δεδομένων της βάσης, με εγκαταστάσεις αποθήκευσης δεδομένων και εγκαταστάσεις Data Mining. Μια αποθήκη δεδομένων ενσωματώνει δεδομένα προερχόμενα από πολλαπλές πηγές και διάφορα χρονικά πλαίσια. Συγκεντρώνει δεδομένα σε ένα πολυδιάστατο χώρο για να σχηματίσουν μερικώς υλοποιημένους «κύβους δεδομένων», **[10], [15]**.

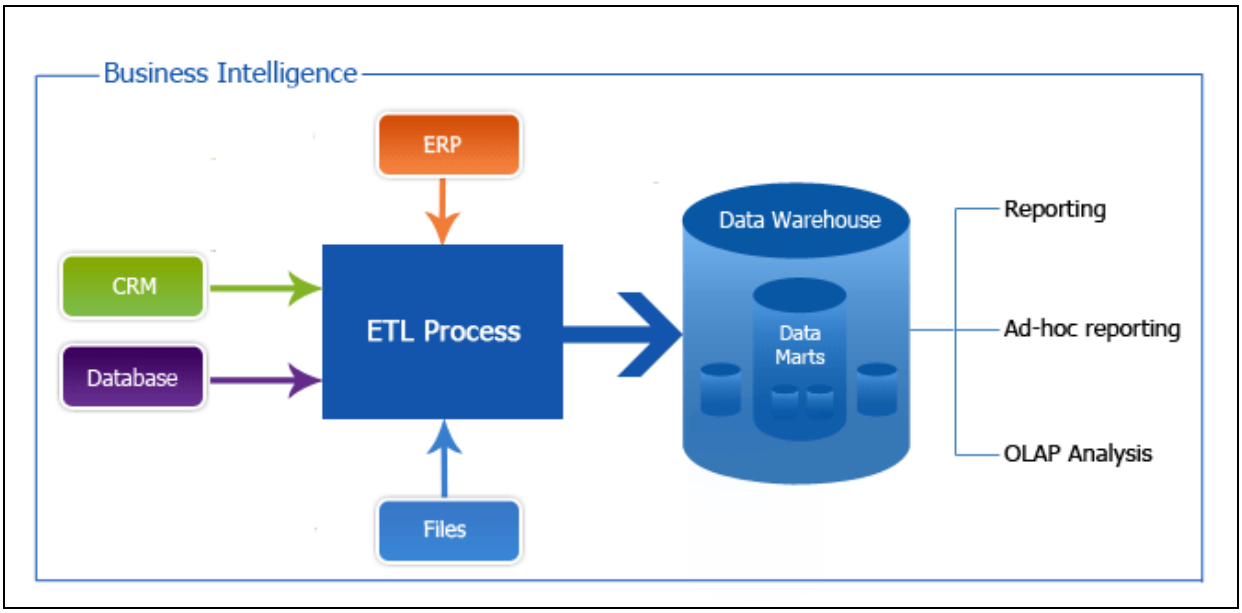
2.5. Εφαρμογές Αξιοποίησης του Data Mining

Η εξόρυξη δεδομένων έχει σημειώσει μεγάλες επιτυχίες σε πολλούς τομείς εφαρμογών. Είναι αδύνατο να απαριθμηθούν όλες οι εφαρμογές όπου η εξόρυξη δεδομένων παίζει κρίσιμο ρόλο. Ανάλυση των τρόπων αξιοποίησης του Data Mining σε τομείς εφαρμογών όπως η βιοπληροφορική και η τεχνολογία λογισμικού, απαιτούν πιο εις βάθος επεξεργασία και είναι πέρα από τις δυνατότητες αυτής της εργασίας. Για να

αποδειχθεί η σημασία της εξόρυξης δεδομένων ως μιας σημαντικής διάστασης στην έρευνα και ανάπτυξη των εφαρμογών , συζητούνται εν συντομία δύο εξαιρετικά επιτυχημένα και δημοφιλή παραδείγματα εφαρμογής Data Mining: η επιχειρηματική ευφυΐα (Business Intelligence) και οι μηχανές αναζήτησης παγκόσμιου ιστού (web search engines).

2.5.1. Επιχειρηματική Ευφυΐα (Business Intelligence)

Είναι ζωτικής σημασίας για τις επιχειρήσεις να αποκτήσουν καλύτερη κατανόηση του εμπορικού πλαισίου της οργάνωσής τους, όπως οι πελάτες τους, η αγορά, η προσφορά και οι πόροι και οι ανταγωνιστές. Οι τεχνολογίες επιχειρηματικής ευφυΐας (Business Intelligence - BI) παρέχουν την ιστορική και την τρέχουσα εικόνα αλλά και προβλέψεις για το μέλλον, για συγκεκριμένες επιχειρηματικές δραστηριότητες όπως βλέπουμε σχηματικά και στην συνέχεια (Εικόνα 2.5). Παραδείγματα που χρησιμοποιείται το BI είναι η ηλεκτρονική αναλυτική επεξεργασία, διαχείριση επιχειρηματικών επιδόσεων, ανταγωνιστική νοημοσύνη, συγκριτική αξιολόγηση, προγνωστικά αναλύσεων κτλ. Πόσο σημαντικό είναι το BI; Χωρίς το Data Mining, πολλές επιχειρήσεις μπορεί να μην ήταν σε θέση να εκτελέσουν αποτελεσματική ανάλυση της αγοράς, να συγκρίνουν τα σχόλια των πελατών σχετικά με παρόμοια προϊόντα, να ανακαλύψουν τα δυνατά σημεία και τις αδυναμίες των ανταγωνιστών τους, να διατηρήσουν υψηλό επίπεδο για τους σημαντικούς πελάτες και να παίρνουν έξυπνες επιχειρηματικές αποφάσεις. Σαφώς, το Data Mining είναι ο πυρήνας του BI. Hon-line αναλυτική επεξεργασία των δεδομένων μέσω των εργαλείων του BI βασίζεται στην αποθήκευση δεδομένων και στην εξόρυξη πολυδιάστατων δεδομένων. Οι τεχνικές ταξινόμησης (classification) και πρόβλεψης (prediction) αποτελούν τον πυρήνα των προγνωστικών αναλύσεων στο Business BI, με πολλές εφαρμογές στην ανάλυση αγορών, τις προμήθειες και τις πωλήσεις. Επιπλέον, η ομαδοποίηση (clustering) διαδραματίζει κεντρικό ρόλο στη σχέση εταιρίας- πελατών, καθώς ομαδοποιεί τους πελάτες με βάση τις ομοιότητές τους. Χρησιμοποιώντας τεχνικές εξόρυξης, μπορούν να γίνουν καλύτερα κατανοητά τα χαρακτηριστικά κάθε ομάδας πελατών και να αναπτυχθούν, π.χ., προσαρμοσμένα προγράμματα ανταμοιβής πελατών, **[7], [10], [15]**.



Εικόνα.2.5: Δομή του Business Intelligence

(Πηγή. http://www.colizeeumtechnology.com/Technologies/DataWarehouse.aspx)

2.5.2. Μηχανές Αναζήτησης Παγκόσμιου Ιστού (Web Search Engines)

Μια μηχανή αναζήτησης παγκόσμιου ιστού (Web Search Engine) είναι ένας εξειδικευμένος υπολογιστής που αναζητά πληροφορίες σε όλο το εύρος του Παγκόσμιου Ιστού, λειτουργώντας πάνω από το Διαδίκτυο. Τα αποτελέσματα αναζήτησης ενός ερωτήματος του χρήστη συχνά επιστρέφονται ως λίστα. Οι επισκέψεις μπορούν να αποτελούνται από ιστοσελίδες, εικόνες και άλλους τύπους αρχείων. Μερικές μηχανές αναζήτησης αναζητούν και επιστρέφουν δεδομένα διαθέσιμα σε δημόσιες βάσεις δεδομένων ή σε ανοιχτούς καταλόγους.

Οι μηχανές αναζήτησης διαφέρουν από τους καταλόγους ιστού διότι οι κατάλογοι ιστού συντηρούνται από ανθρώπους-χρήστες ενώ οι μηχανές αναζήτησης λειτουργούν αλγοριθμικά ή με μείγμα αλγοριθμικής και ανθρώπινης εισόδου. Οι μηχανές αναζήτησης είναι ουσιαστικά πολύ μεγάλες εφαρμογές Data Mining. Σε διάφορα είδη δεδομένων, οι τεχνικές εξόρυξης χρησιμοποιούνται σε όλες τις πτυχές των μηχανών αναζήτησης, που κυμαίνονται από την ανίχνευση (π.χ. να αποφασίζει ποιες σελίδες πρέπει να ανιχνεύονται και τις συχνότητες ανίχνευσής τους), ευρετηρίαση (π.χ. επιλογή των σελίδων που θα ευρετηριαστούν και αποφασίζει σε ποιο βαθμό θα πρέπει να είναι ο δείκτης κατασκευής)

και την αναζήτηση (π.χ. να αποφασιστεί ο τρόπος ταξινόμησης των σελίδων, ποιες διαφημίσεις πρέπει να προστεθούν και πώς μπορούν να εξατομικευθούν).

Οι μηχανές αναζήτησης θέτουν μεγάλες προκλήσεις στο ίδιο το Data Mining. Πρώτον, πρέπει να χειριστούν μια τεράστια και συνεχώς αυξανόμενη ποσότητα δεδομένων. Συνήθως, τα δεδομένα αυτά δεν μπορούν να υποστούν επεξεργασία ή υπάρχουν λίγα διαθέσιμα για την εργασία αυτή φυσικά υπολογιστικά μηχανήματα. Αντ' αυτού, οι μηχανές αναζήτησης πρέπει συχνά να χρησιμοποιούν τα υπολογιστικά νέφη (Clouds), τα οποία αποτελούνται από χιλιάδες ή και εκατοντάδες χιλιάδες υπολογιστές που συνεργάζονται με το τεράστιο όγκο των δεδομένων. Η κλιμάκωση προς τα πάνω των μεθόδων εξόρυξης δεδομένων ώστε να λειτουργούν σε νέφη (Clouds) υπολογιστών και μεγάλα κατανεμημένα σύνολα δεδομένων είναι ένας τομέας που θα απασχολήσει σοβαρά την έρευνα για τα επόμενα 15 τουλάχιστον χρόνια.

Δεύτερον, οι μηχανές αναζήτησης πρέπει συχνά να ασχολούνται με ηλεκτρονικά δεδομένα. Μια μηχανή αναζήτησης μπορεί να είναι σε θέση να προσφέρει την κατασκευή ενός μοντέλου εκτός σύνδεσης σε τεράστια σύνολα δεδομένων. Για να γίνει αυτό, πρέπει να κατασκευαστεί ένας ταξινομητής ερωτήματος ο οποίος θα εκχωρεί ένα ερώτημα αναζήτησης σε προκαθορισμένες κατηγορίες, με βάση το θέμα του ερωτήματος (δηλ. αν το ερώτημα αναζήτησης "μήλο" (apple) προορίζεται για την ανάκτηση πληροφοριών για ένα φρούτο ή μια μάρκα υπολογιστών). Εάν ένα μοντέλο είναι κατασκευασμένο εκτός σύνδεσης, η εφαρμογή του ηλεκτρονικού μοντέλου πρέπει να είναι αρκετά γρήγορη ώστε να απαντά σε ερωτήματα χρηστών σε πραγματικό χρόνο.

Μια άλλη πρόκληση είναι η συντήρηση και η σταδιακή ενημέρωση ενός μοντέλου για τη γρήγορη ανάπτυξη ροών δεδομένων. Για παράδειγμα, ένας ταξινομητής ερωτήματος μπορεί να χρειαστεί να επικαιροποιεί συνεχώς τις πληροφορίες, καθώς τα νέα ερωτήματα φέρνουν και νέο-αναδυόμενες και προκαθορισμένες κατηγορίες και η κατανομή των δεδομένων μπορεί να αλλάξει. Οι περισσότερες από τις υπάρχουσες μεθόδους κατάρτισης μοντέλων είναι off-line και συνεπώς δεν μπορούν να χρησιμοποιηθούν σε ένα τέτοιο σενάριο.

Τρίτον, οι μηχανές αναζήτησης συχνά έχουν να αντιμετωπίσουν ερωτήματα που ζητούνται μόνο με πολύ μικρή συχνότητα. Ας υποτεθεί ότι μια μηχανή αναζήτησης θέλει να απαντήσει ένα ερώτημα σχετικά με το περιβάλλον. Όταν ένας χρήστης θέτει το ερώτημα, η μηχανή αναζήτησης προσπαθεί να συμπεράνει το πλαίσιο του ερωτήματος. Για το σκοπό

αυτό χρησιμοποιεί το προφίλ του χρήστη και το ιστορικό των ερωτημάτων του, για να επιστρέψει πιο προσαρμοσμένες απαντήσεις μέσα σε ένα μικρό κλάσμα του δευτερολέπτου. Αυτό κάνουν τα γνωστά «Cookies» στις σελίδες του παγκόσμιου ιστού. Ωστόσο, αν και το ο συνολικός αριθμός ερωτημάτων που αναζητούνται μπορεί να είναι τεράστιος, τα περισσότερα ερωτήματα μπορεί να ζητηθούν μόνο μία φορά ή μόνο μερικές (λίγες) φορές, **[7], [10], [15]**.

2.6. Βασικά πρότυπα επαναλαμβανόμενης αναζήτησης στην εξόρυξη δεδομένων

Η εξόρυξη δεδομένων καταλήγει να αναζητά επαναλαμβανόμενες σχέσεις («πρότυπα») σε ένα συγκεκριμένο σύνολο δεδομένων. Στη συνέχεια παρουσιάζονται οι πλέον βασικές έννοιες της συχνής εξόρυξης προτύπων για την ανακάλυψη ενδιαφερουσών συσχετίσεων μεταξύ αντικειμένων σε σχεσιακές βάσεων δεδομένων. Ένα παράδειγμα είναι η ανάλυση του «καλαθιού αγοράς» στο πλαίσιο του e-shopping, που είναι μια πρώιμη μορφή συχνής εξόρυξης προτύπων.

2.6.1. Ανάλυση Καλαθιού Αγοράς

Η συχνή εξόρυξη αντικειμένων οδηγεί στην ανακάλυψη συσχετίσεων μεταξύ αντικειμένων σε μεγάλα σύνολα δεδομένων, κυρίως στις σχεσιακές βάσεις δεδομένων. Τεράστιες ποσότητες δεδομένων συνεχώς συλλέγονται και αποθηκεύονται, ενώ πλέον έχει τεράστιο ενδιαφέρον για πολλούς φορείς, ιδιωτικούς και δημόσιους, η εξόρυξη αυτών των δεδομένων από τις βάσεις δεδομένων τους (από τεράστιες πολυεθνικές εταιρίες έως νοσοκομειακές μονάδες, κτλ). Η ανακάλυψη ενδιαφερουσών σχέσεων συσχέτισης ανάμεσα σε τεράστιες ποσότητες συναλλαγών μπορεί να βοηθήσει πολλές επιχειρήσεις στις διαδικασίες λήψης αποφάσεων όπως ο σχεδιασμός δράσεων, η διασταύρωση πελατών, η ανάλυση συμπεριφοράς αγορών, κτλ.

Ένα χαρακτηριστικό παράδειγμα συχνής εξόρυξης αντικειμένων είναι η ανάλυση του καλαθιού αγοράς στο πλαίσιο των διαδικτυακών αγορών αγαθών. Αυτή η διαδικασία αναλύει τις καταναλωτικές συνήθειες των πελατών εντοπίζοντας συσχετισμούς μεταξύ των διαφόρων ειδών που οι πελάτες τοποθετούν στα "καλάθια αγορών". Η ανακάλυψη αυτών των σχέσεων μπορεί να βοηθήσει στην ανάπτυξη στρατηγικών μάρκετινγκ, προσφέροντας

διορατικότητα ως προς το ποια στοιχεία καταναλώνονται ή επιλέγονται πιο συχνά από τους πελάτες. Για παράδειγμα, εάν οι πελάτες αγοράζουν γάλα, πόσο πιθανό είναι να αγοράσουν μαζί και ψωμί (και ποιο είδος ψωμιού); Σε αγορές ή ενοικιάσεις βιβλίων σε βιβλιοθήκες και καταστήματα εμπορίας βιβλίων, η αγορά ενοικίαση ενός συγκεκριμένου τίτλου βιβλίου πόσο συχνά συνοδεύεται από αγορά ή ενοικίαση και συγκεκριμένου άλλου τίτλου; Αυτές οι πληροφορίες μπορούν να οδηγήσουν σε αύξηση των πωλήσεων ή, σε περιπτώσεις δανεισμών, σε αύξηση των διαθέσιμων τεμαχίων των προϊόντων προς δανεισμό (εάν ένα βιβλίο π.χ. έχει έντονη ζήτηση) βοηθώντας το επιλεκτικό μάρκετινγκ ή ακόμα και το σχεδιασμό της προθήκης (φυσικής ή ηλεκτρονικής) της εταιρίας, **[17], [18], [22]**.

2.7. Data Mining για την Επιστήμη και τους Μηχανικούς

Στο παρελθόν, ήταν σύνηθες οι αλγόριθμοι ανάλυσης επιστημονικών δεδομένων να χειρίζονται σχετικά μικρά και ομοιογενή σύνολα δεδομένων. Τέτοια τυπικά δεδομένα αναλύθηκαν με τη χρήση μιας «διαμόρφωσης όπου δημιουργεί ένα μοντέλο και αξιολογεί το παράδειγμά του». Σε αυτές τις περιπτώσεις, υπάρχουν τεχνικές στατιστικής που χρησιμοποιούνται συνήθως για την ανάλυσή τους.

Η μαζική συλλογή δεδομένων και οι τεχνολογίες αποθήκευσης έχουν αλλάξει πρόσφατα το τοπίο της ανάλυσης επιστημονικών δεδομένων. Σήμερα, τα επιστημονικά δεδομένα μπορούν να συγκεντρωθούν με πολύ υψηλότερες ταχύτητες και χαμηλότερο κόστος. Αυτό έχει ως αποτέλεσμα τη συσσώρευση τεράστιων όγκων δεδομένων πολλών διαστάσεων, δεδομένων ροής και ετερογενών δεδομένα που περιέχουν πλούσιες πληροφορίες.

Κατά συνέπεια, οι επιστημονικές εφαρμογές μετατοπίζονται από το "υποθετικό και δοκιμαστικό" πρότυπο προς "συλλογή και αποθήκευση δεδομένων’. Αυτή η μετατόπιση δημιουργεί νέες προκλήσεις για την εξόρυξη δεδομένων. Μεγάλοι όγκοι δεδομένων συλλέχθηκαν από επιστημονικούς τομείς (αστρονομία, μετεωρολογία, γεωλογία και βιολογικές επιστήμες) χρησιμοποιώντας τηλεσκόπια, πολυφασματικούς απομακρυσμένους δορυφορικούς αισθητήρες υψηλής ανάλυσης, και νέες γενεές τεχνολογιών συλλογής και ανάλυσης βιολογικών δεδομένων, **[7], [12]**.

Μεγάλα σύνολα δεδομένων παράγονται επίσης λόγω γρήγορων αριθμητικών προσομοιώσεων σε διάφορους τομείς, όπως η μοντελοποίηση του κλίματος και των οικοσυστημάτων, η χημική μηχανική, η δυναμική των υγρών και η δομική μηχανική. Στη συνέχεια παρατίθενται μερικές από τις προκλήσεις που προκάλεσαν οι αναδυόμενες επιστημονικές εφαρμογές της εξόρυξης δεδομένων.

✓ **Αποθήκες Δεδομένων και Προεπεξεργασία Δεδομένων**: Η Προεπεξεργασία δεδομένων (Data (Pre-)Processing) και οι αποθήκες δεδομένων (Data Warehouse) είναι ζωτικής σημασίας για την εξόρυξη δεδομένων. Συχνά μια αποθήκη απαιτεί εύρεση μεθόδων για την αντιμετώπιση ασυνεπών ή αντιφατικών δεδομένων που συλλέγονται σε διαφορετικά περιβάλλοντα και σε διαφορετικές χρονικές περιόδους. Αυτό με τη σειρά του απαιτεί συστήματα αναφοράς, γεωμετρία, μετρήσεις και ακρίβεια. Απαιτούνται μέθοδοι για την ενσωμάτωση δεδομένων από ετερογενείς πηγές και για προσδιορισμό των γεγονότων. Παράδειγμα αποτελούν τα δεδομένα για το κλίμα και το οικοσύστημα, τα οποία είναι χωρικά και χρονικά και απαιτούν διασταυρούμενες γεω-περιβαλλοντικές πληροφορίες. Ένα σημαντικό πρόβλημα στην ανάλυση τέτοιων δεδομένων είναι ότι υπάρχουν πάρα πολλά γεγονότα στον χωρικό άξονα, αλλά πολύ λίγα στον χρονικό άξονα. Για παράδειγμα, τα γεγονότα του καταστροφικού φαινομένου «El-Nino» συμβαίνουν μόνο κάθε τέσσερα έως επτά χρόνια και τα προηγούμενα δεδομένα σχετικά με αυτά ενδέχεται να μην έχουν συλλεχθεί τόσο συστηματικά όσο τα νεώτερα, **[10], [15]**.

✓ **Εξόρυξη Πολύπλοκων Τύπων Δεδομένων**: Τα επιστημονικά σύνολα δεδομένων είναι συνήθως ανομοιογενή. Αυτά συνήθως περιλαμβάνουν ημι-δομημένα και αδόμητα δεδομένα, όπως δεδομένα πολυμέσων και δεδομένα ροής, καθώς και δεδομένα με κρυμμένη σημασία (π.χ. γονιδιωματικά δεδομένα στη βιολογία). Σταθερές μέθοδοι ανάλυσης είναι απαραίτητες για το χειρισμό χωροχρονικών δεδομένων, βιολογικών δεδομένων και σύνθετων σημασιολογικών σχέσεων. Για παράδειγμα, στη βιοπληροφορική, ένα ερευνητικό πρόβλημα είναι να προσδιοριστούν οι ρυθμιστικές επιδράσεις στα γονίδια. Η ρύθμιση γονιδίων αναφέρεται στο πώς ενεργοποιούνται (ή απενεργοποιούνται) τα γονίδια σε ένα κύτταρο για τον προσδιορισμό των λειτουργιών του κυττάρου. Οι διαφορετικές βιολογικές διεργασίες περιλαμβάνουν διαφορετικά σύνολα γονιδίων που δρουν μαζί σε επακριβώς ρυθμιζόμενα πρότυπα. Έτσι, για να γίνει κατανοητή μια βιολογική διαδικασία πρέπει να εντοπιστούν τα συμμετέχοντα γονίδια και οι ρυθμιστικές αρχές τους. Αυτό απαιτεί την ανάπτυξη των εξελιγμένων μεθόδων εξόρυξης δεδομένων για

την ανάλυση μεγάλων σειρών βιολογικών δεδομένων σε αναζήτηση ενδείξεων σχετικά με τις ρυθμιστικές επιδράσεις σε συγκεκριμένα γονίδια, με την εύρεση τμημάτων DNA, **[10], [15]**.

- ✓ **Γραμμική και Βασιζόμενη στο Δίκτυο Εξόρυξη**: Είναι συχνά δύσκολη ή αδύνατη η μοντελοποίηση ορισμένων φυσικών φαινομένων και διαδικασιών εξαιτίας περιορισμών των υφιστάμενων προσεγγίσεων μοντελοποίησης. Εναλλακτικά, μπορούν να χρησιμοποιηθούν γραφήματα και δίκτυα για να εντοπιστούν πολλά από τα χωρικά, τοπολογικά, γεωμετρικά, βιολογικά και άλλα σχεσιακά χαρακτηριστικά που υπάρχουν σε σύνολα επιστημονικών δεδομένων. Στη γραφική παράσταση ή στη μοντελοποίηση δικτύου, το αντικείμενο-στόχος της εξόρυξης αντιπροσωπεύεται από μια κορυφή σε ένα γράφημα και τις άκρες μεταξύ των κορυφών αντιπροσωπεύουν σχέσεις μεταξύ αντικειμένων. Για παράδειγμα, γραφήματα χρησιμοποιούνται για να παραστήσουν μοντέλα για χημικές δομές, βιολογικές οδούς ή δεδομένα που παράγονται από αριθμητικές προσομοιώσεις, όπως προσομοιώσεις ροής ρευστού. Η επιτυχία της διαμόρφωσης γραφημάτων ή δικτύων εξαρτάται από τις βελτιώσεις στην επεκτασιμότητα και την αποδοτικότητα πολλών γραφικών εργασιών εξόρυξης δεδομένων όπως η ταξινόμηση, η συχνή εξόρυξη προτύπων και η συσπείρωση των δεδομένων, **[10], [15]**.

- ✓ **Εργαλεία Απεικόνισης**: Ορισμένες κατηγορίες χρηστών θέτουν υψηλού επιπέδου απαιτήσεις ως προς τις διασυνδέσεις και τα εργαλεία οπτικοποίησης για τα συστήματα εξόρυξης επιστημονικών δεδομένων. Αυτά θα πρέπει να ενσωματωθούν σε υπάρχοντα συστήματα δεδομένων και πληροφορικής και για το συγκεκριμένο επιστημονικό τομέα, τα οποία και θα καθοδηγήσουν τους ερευνητές και τους απλούς χρήστες στην αναζήτηση μοντέλων, την απεικόνιση των ανακαλυφθέντων μοτίβων και τη χρήση τους για την ανακάλυψη γνώσης στην κατασκευή τους.

Η εξόρυξη δεδομένων στην τεχνολογία μοιράζεται πολλές ομοιότητες με την εξόρυξη δεδομένων στην επιστήμη. Και οι δύο πρακτικές συχνά συλλέγουν τεράστιες ποσότητες δεδομένων και απαιτούν την προεπεξεργασία αυτών των δεδομένων, την αποθήκευση και την επεκτάσιμη εξόρυξη πολύπλοκων τύπων δεδομένων. Και οι δύο χρησιμοποιούν συνήθως ποιοτική απεικόνιση και καλή χρήση γραφημάτων και δικτύων. Επιπλέον, πολλοί μηχανικοί χρειάζονται απαντήσεις ώστε να δράσουν ή να αντιδράσουν σε πραγματικό χρόνο και έτσι οι ροές δεδομένων εξορύσσονται σε πραγματικό χρόνο, με αποτέλεσμα η διάσταση του χρόνου να μετατρέπεται σε ένα κρίσιμο στοιχείο, **[7], [15]**.

Μεγάλες ποσότητες δεδομένων που προκύπτουν από την ανθρώπινη επικοινωνία εμφανίζονται στην καθημερινότητά μας. Η επικοινωνία υπάρχει σε πολλές μορφές, συμπεριλαμβανομένων ειδήσεων, blogs, άρθρων, ιστοσελίδων, συζητήσεων σε απευθείας σύνδεση, κριτικών για προϊόντα, δημοσιεύσεων κειμένου τύπου twitter, μηνυμάτων κειμένου ή/και πολυμέσων, διαφημίσεων, τόσο στον Παγκόσμιο Ιστό όσο και ειδικότερα εντός διάφορων τύπων κοινωνικών δικτύων. Ως εκ τούτου, η εξόρυξη δεδομένων στο χώρο των κοινωνικών ανταλλαγών ή στο χώρο των επιστημονικών ερευνών έχει γίνει όλο και πιο δημοφιλής. Τα αποτελέσματα της ανάλυσης μπορούν να χρησιμοποιούνται για την πρόβλεψη των τάσεων, τη βελτίωση της εργασίας και τη βοήθεια στη λήψη αποφάσεων.

Η επιστήμη των υπολογιστών, ειδικότερα, παράγει συγκεκριμένους τύπους δεδομένων. Για παράδειγμα, τα προγράμματα λογισμικού των υπολογιστών μπορεί να είναι μεγάλα, και η εκτέλεση τους συχνά να δημιουργεί ίχνη μεγάλου μεγέθους. Τα Δίκτυα Υπολογιστών μπορεί να έχουν σύνθετες δομές και οι ροές του δικτύου μπορεί να είναι δυναμικές και μαζικές. Τα Δίκτυα Αισθητήρων δημιουργούν μεγάλες ποσότητες δεδομένων από μετρήσεις που λαμβάνουν οι αισθητήρες, με ποικίλη αξιοπιστία.

Αυτά τα συγκεκριμένα είδη δεδομένων παρέχουν εύφορο έδαφος για την εξόρυξη δεδομένων. Η εξόρυξη δεδομένων στην επιστήμη των υπολογιστών μπορεί να χρησιμοποιηθεί για να βοηθήσει στην παρακολούθηση της κατάστασης του υπολογιστικού συστήματος, τη βελτίωση της απόδοσης του συστήματος, τον εντοπισμό σφαλμάτων λογισμικού, την ανίχνευση λογοκλοπής, την ανάλυση σφαλμάτων του συστήματος υπολογιστών ή την αποκάλυψη εισβολών το δίκτυο και την αναγνώριση δυσλειτουργιών του συστήματος.

Η εξόρυξη δεδομένων είτε για το λογισμικό είτε για το υλικό του υπολογιστικού συστήματος μπορεί να λειτουργήσει στατικά ή δυναμικά (δηλ. δεδομένα που βασίζονται σε ροή), ανάλογα με το εάν το σύστημα αφαιρεί εκ των προτέρων τα ίχνη μετά την ανάλυση ή εάν πρέπει να αντιδράσει σε πραγματικό χρόνο για τη διαχείριση των ηλεκτρονικών δεδομένων. Στο πεδίο αυτό έχουν αναπτυχθεί διάφορες μέθοδοι, οι οποίες ενσωματώνουν και επεκτείνουν μεθόδους από το χώρο της μηχανικής μάθησης, της εξόρυξης δεδομένων, της μηχανικής λογισμικού, και της αναγνώρισης συστημάτων. Η εξόρυξη δεδομένων στην επιστήμη των υπολογιστών είναι ένας ενεργός και πλούσιος τομέας για τους ερευνητές λόγω των μοναδικών προκλήσεων που προσφέρει, **[10]**.

2.8. Προβλήματα στο Data Mining

Η εξόρυξη δεδομένων είναι ένας δυναμικός και ταχέως αναπτυσσόμενος τομέας με μεγάλες δυνατότητες. Κλείνοντας την ανάλυση στο πλαίσιο της παρούσας εργασίας, αναφέρονται εν συντομία τα κυριότερα ζητήματα ή σημεία προβληματισμού της έρευνας και των ερευνητών περί το Data Mining, τα οποία χωρίζονται σε πέντε ομάδες:

- ✓ Μεθοδολογία εξόρυξης
- ✓ Αλληλεπίδραση χρηστών
- ✓ Αποδοτικότητα και κλιμάκωση
- ✓ Ποικιλομορφία στους τύπους δεδομένων
- ✓ Την εξόρυξη δεδομένων και την κοινωνία.

Πολλά από αυτά τα ζητήματα έχουν ήδη αντιμετωπιστεί σε κάποιο βαθμό από την πρόσφατη έρευνα και την ανάπτυξη για το Data Mining και τώρα εξετάζονται οι απαιτήσεις του Data Mining. Πολλοί φορείς βρίσκονται ακόμη στα αρχικά στάδια της έρευνας στα ερωτήματα αυτά. Τα προβλήματα που διαρκώς προκύπτουν διατηρούν και τονώνουν την έρευνα και περαιτέρω διερεύνησης και βελτίωσης της εξόρυξης δεδομένων.

Εισαγωγή

Πώς μπορούν να παρουσιαστούν και να αποδοθούν αποτελεσματικά στον χρήστη τα δεδομένα; Η οπτικοποίηση δεδομένων (data visualization) στοχεύει στην παρουσίαση των δεδομένων στο χρήστη με σαφώς αποτελεσματικό για την πρόσληψή τους τρόπο, μέσω γραφικών αναπαραστάσεων διαφόρων τύπων. Η οπτικοποίηση δεδομένων χρησιμοποιείται εκτεταμένα σε πολλές εφαρμογές όπως για παράδειγμα στην εργασία, τη διαχείριση επιχειρηματικής δραστηριότητας, την παρακολούθηση της προόδου των εργασιών, κτλ. Το μεγάλο στοίχημα του Data Visualization είναι να εφαρμοστούν τεχνικές οπτικοποίησης για να ανακαλυφθούν σχέσεις δεδομένων που διαφορετικά δεν θα ήταν είναι εύκολα παρατηρήσιμες με την απλή εποπτική εξέταση των ακατέργαστων δεδομένων.

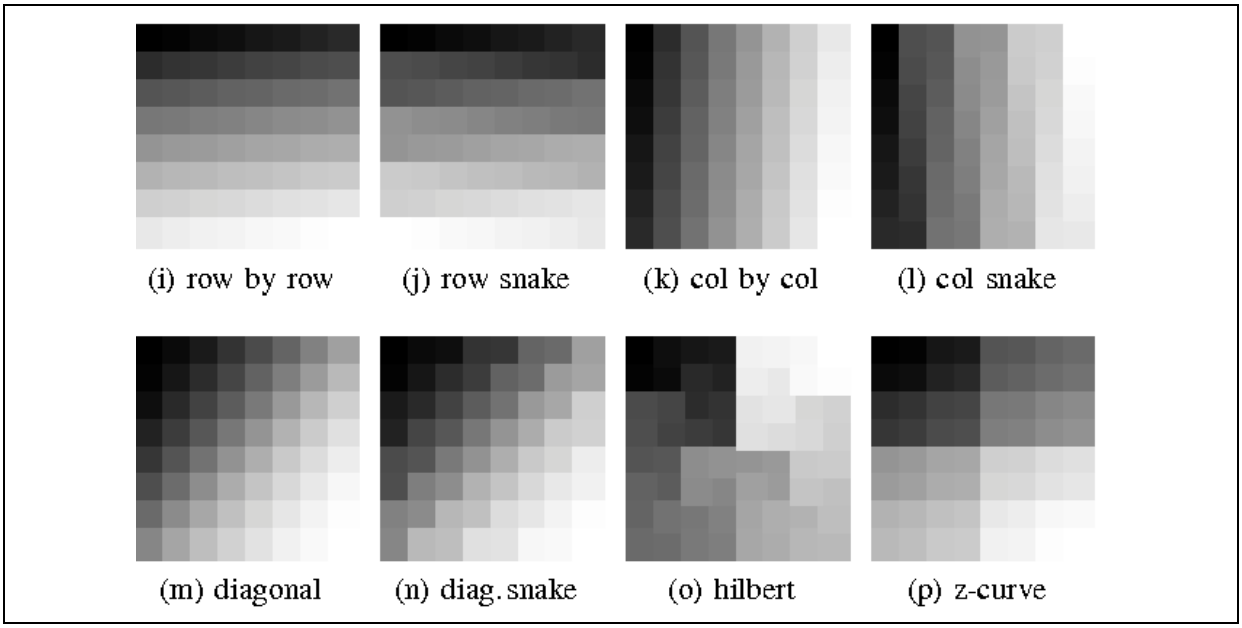
Σήμερα χρησιμοποιείται επίσης η οπτικοποίηση δεδομένων για να δημιουργηθούν ελκυστικά, διασκεδαστικά και ενδιαφέροντα γραφικά. Σε αυτή την ενότητα, παρουσιάζονται εν συντομία οι βασικές έννοιες της οπτικοποίησης δεδομένων. Αρχή γίνεται με τα πολυδιάστατα δεδομένα όπως αυτά που αποθηκεύονται σε σχεσιακές βάσεις δεδομένων. Εξετάζονται διάφορες αντιπροσωπευτικές προσεγγίσεις, συμπεριλαμβανομένων τεχνικών με εικονοστοιχεία, τεχνικές γεωμετρικής προβολής, τεχνικές βασισμένες σε εικονίδια και ιεραρχικές τεχνικές, καθώς και τεχνικές που βασίζονται σε γραφήματα.

3.1. Τεχνικές Οπτικοποίησης με Εικονοστοιχεία

Ένας απλός τρόπος για την απεικόνιση της αξίας μιας διάστασης ή ενός φυσικού μεγέθους είναι να χρησιμοποιηθεί ένα εικονοστοιχείο (pixel), οπότε το χρώμα του εικονοστοιχείου αντικατοπτρίζει την τιμή του φυσικού μεγέθους. Για ένα σύνολο δεδομένων **m** διαστάσεων, που αναπαρίσταται με εικονοστοιχεία, οι τεχνικές δημιουργούν **m** παράθυρα στην οθόνη, ένα για κάθε διάσταση. Για τη διάσταση **m** οι τιμές μιας εγγραφής χαρτογραφούνται με εικονοστοιχεία στις αντίστοιχες θέσεις στα παράθυρα. Τα

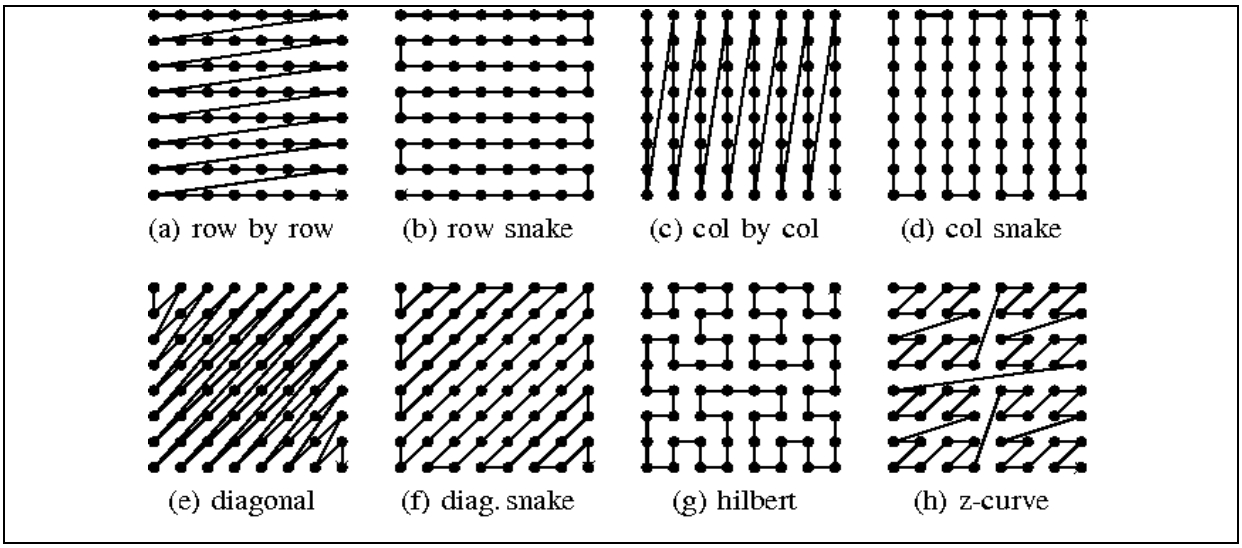
χρώματα των εικονοστοιχείων αντικατοπτρίζουν τις αντίστοιχες τιμές. Μέσα σε ένα παράθυρο, οι τιμές των δεδομένων είναι διατεταγμένες σε μία κοινή για όλα τα παράθυρα σειρά. Η ταξινόμηση μπορεί να επιτευχθεί με τη διαλογή όλων των αρχείων δεδομένων με τέτοιο τρόπο ώστε να έχει νόημα για το συγκεκριμένο αποτέλεσμα-στόχο (Εικόνα 3.1).

Ας υποτεθεί ότι μια εταιρεία διατηρεί έναν πίνακα πληροφοριών πελατών, με τέσσερις διαστάσεις: το εισόδημα, το πιστωτικό όριο, τον όγκο συναλλαγών και την ηλικία του κάθε πελάτη. Μπορεί να αναλυθεί η σχέση μεταξύ του εισοδήματος και των άλλων τριών χαρακτηριστικών με οπτικοποίηση. Επίσης μπορούν να ταξινομηθούν όλοι οι πελάτες σε αύξουσα σειρά εισοδήματος και να χρησιμοποιηθεί αυτή η σειρά για να διευθετηθούν τα δεδομένα πελατών σε πολλά παράθυρα απεικόνισης. Σε επίπεδο εικονοστοιχείου, τα χρώματα επιλέγονται έτσι ώστε όσο μικρότερη είναι η τιμή του μεγέθους, τόσο πιο ελαφριά να είναι η σκίαση. Χρησιμοποιώντας οπτικοποίηση εικονοστοιχείων, μπορεί εύκολα να διαπιστωθεί ότι το πιστωτικό όριο αυξάνεται όσο το εισόδημα αυξάνει ή ότι οι πελάτες των οποίων το εισόδημα βρίσκεται στο μέσο όρο είναι πιο πιθανό να αγοράσουν περισσότερα από την εταιρεία, καθώς και πάρα πολλούς άλλους ανάλογου τύπου συσχετισμούς μεταξύ των δεδομένων, **[13], [14]**.



Εικόνα.3.1: Οπτικοποίηση Δεδομένων με εικονοστοιχεία σε διάφορες μορφές ταξινόμησης των δεδομένων (Πηγή. https://www.semanticscholar.org/paper/Pixel-Oriented-Visualization-of-Change-in-Social-Stein-Wegener/89781726d303b54b908029cded8229ee7a42770c)

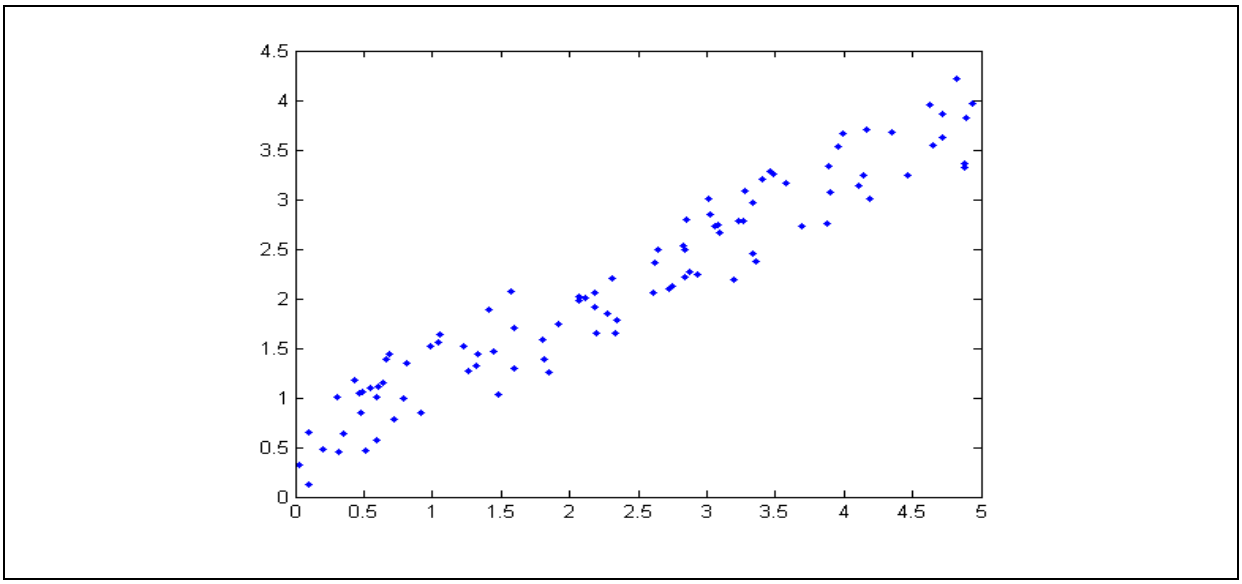
Στις τεχνικές με εικονοστοιχεία, οι εγγραφές δεδομένων μπορούν επίσης να ταξινομηθούν σε σχέση με μια ερώτηση. Με δεδομένο ένα ερώτημα, ταξινομούνται όλα τα αρχεία σε φθίνουσα σειρά σε σχέση με το ερώτημα. Η συμπλήρωση ενός παραθύρου με την τοποθέτηση των αρχείων δεδομένων με γραμμικό τρόπο μπορεί να μην λειτουργήσει καλά για ένα ευρύ παράθυρο. Το πρώτο εικονοστοιχείο στη σειρά είναι πολύ μακριά από το τελευταίο εικονοστοιχείο της προηγούμενης σειράς. Ένας τρόπος για να λυθεί αυτό το πρόβλημα, θα ήταν να διαμορφωθούν τα αρχεία δεδομένων πάνω σε κάποια καμπύλη πλήρωσης του χώρου, όπως οι καμπύλες στην συνέχεια (Εικόνα 3.2), **[14]**.



Εικόνα.3.2: Οπτικοποίηση Δεδομένων με εικονοστοιχεία σε διάφορες μορφές πλήρωσης χώρου (Πηγή. https://www.semanticscholar.org/paper/Pixel-Oriented-Visualization-of-Change-in-Social-Stein-Wegener/89781726d303b54b908029cded8229ee7a42770c)

3.2. Τεχνικές Γεωμετρικής Προβολής

Ένα μειονέκτημα των τεχνικών απεικόνισης με εικονοστοιχεία είναι ότι δεν μπορούν να βοηθήσουν πολύ στην κατανόηση της κατανομής των δεδομένων στον αδιαφανή χώρο. Για παράδειγμα δεν δείχνουν εάν υπάρχει μια πυκνή περιοχή σε ένα πολυδιάστατο υποχώρο. Οι τεχνικές γεωμετρικής προβολής βοηθούν τους χρήστες να βρουν ενδιαφέρουσες προβολές πολυδιάστατων δεδομένων. Η κεντρική πρόκληση που οι τεχνικές γεωμετρικής προβολής προσπαθούν να αντιμετωπίσουν είναι πώς να απεικονιστεί ένας χώρος μεγάλων διαστάσεων σε μια οθόνη μόνο δύο διαστάσεων (2-D), δηλαδή στο επίπεδο.



Εικόνα.3.3: Οπτικοποίηση 2D χρησιμοποιώντας γραφική παράσταση διασποράς

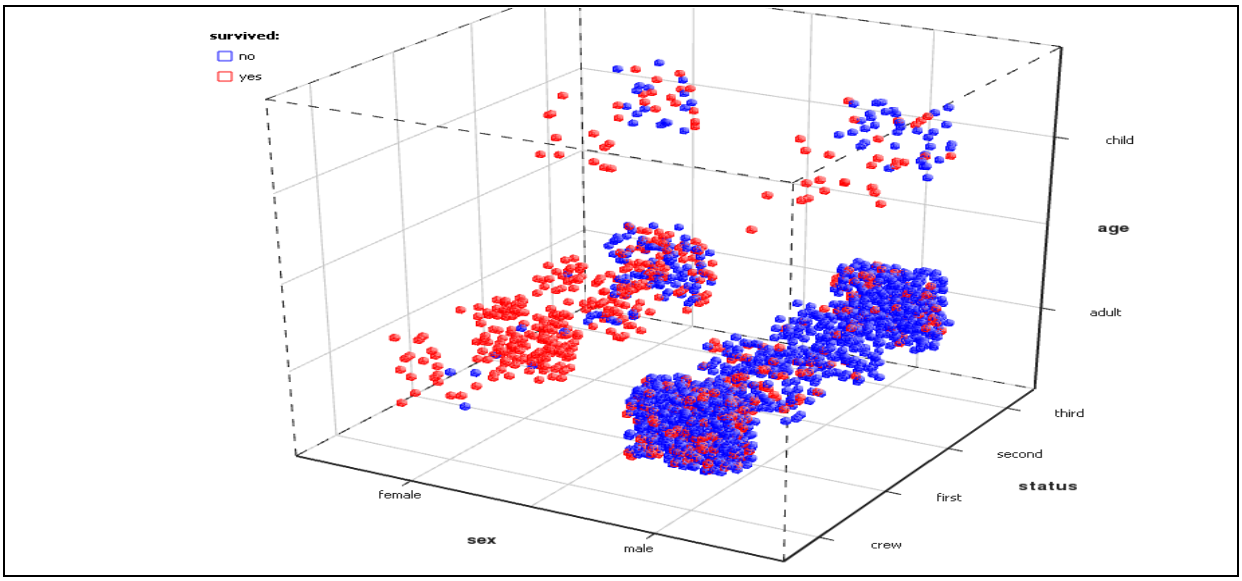
(Πηγή. https://www.xlstat.com/en/solutions/features/scatter-plots)

Μια γραφική παράσταση διασποράς δύο διαστάσεων (2-D scatter plot) εμφανίζει δεδομένα 2-D χρησιμοποιώντας καρτεσιανές συντεταγμένες όπως βλέπουμε καθαρά και στην προηγούμενη εικόνα (Εικόνα 3.3). Μια τρίτη διάσταση θα μπορούσε να προστεθεί χρησιμοποιώντας διαφορετικά χρώματα ή σχήματα για να αντιπροσωπεύει διαφορετικά σημεία δεδομένων. Οι διαστάσεις X και Y θα μπορούσαν να είναι τα δύο χωρικά χαρακτηριστικά και η τρίτη διάσταση (έστω Z) να αντιπροσωπεύεται από διαφορετικά σχήματα.

Ένα διάγραμμα διασποράς τριών διαστάσεων (3-D scatter plot) χρησιμοποιεί τρεις άξονες σε ένα καρτεσιανό σύστημα συντεταγμένων, έστω X, Y και Z. Εάν χρησιμοποιεί επίσης χρώμα, μπορεί να εμφανίσει έως μία $4^{η}$ διάσταση δεδομένων. Για σύνολα δεδομένων με περισσότερες από τέσσερις διαστάσεις, τα διαγράμματα διασποράς είναι συνήθως αναποτελεσματικά.

Η τεχνική γραφικής παράστασης μήτρας διασποράς (scatter matrix) είναι μια χρήσιμη επέκταση του διαγράμματος διασποράς. Για ένα μη διακριτό σύνολο δεδομένων, μια μήτρα διασποράς είναι ένα πλέγμα **n x n** από διαγράμματα διασποράς 2-D που παρέχει μια απεικόνιση (προβολή) της κάθε διάστασης πάνω με κάθε άλλη διάσταση. Για παράδειγμα σε ένα σύνολο δεδομένων ενός Πανεπιστημίου που αποτελείται από περισσότερα από 450 δείγματα (άτομα) από κάθε ένα από τα δυο φύλα φοιτητών. Στο σύνολο των δεδομένων υπάρχουν έξι (6) διαστάσεις: οπτικοποίηση με χρήση διαγραμμάτων διασποράς 3-D και

χρώματος, όπως προαναφέρθηκε, μπορεί να αστικοποιήσει μόνο τις 4 από τις 6 διαστάσεις, π.χ. (Εικόνα 3.4) τα δυο φύλα των φοιτητών, τα τέσσερα έτη σπουδών, δυο κατηγορίες ηλικιών των φοιτητών και μία επιπλέον διάσταση ("survived" – κόκκινο ή μπλε χρώμα). Η μήτρα διασποράς καθίσταται λιγότερο αποτελεσματική καθώς αυξάνονται οι διαστάσεις, **[11], [13]**.

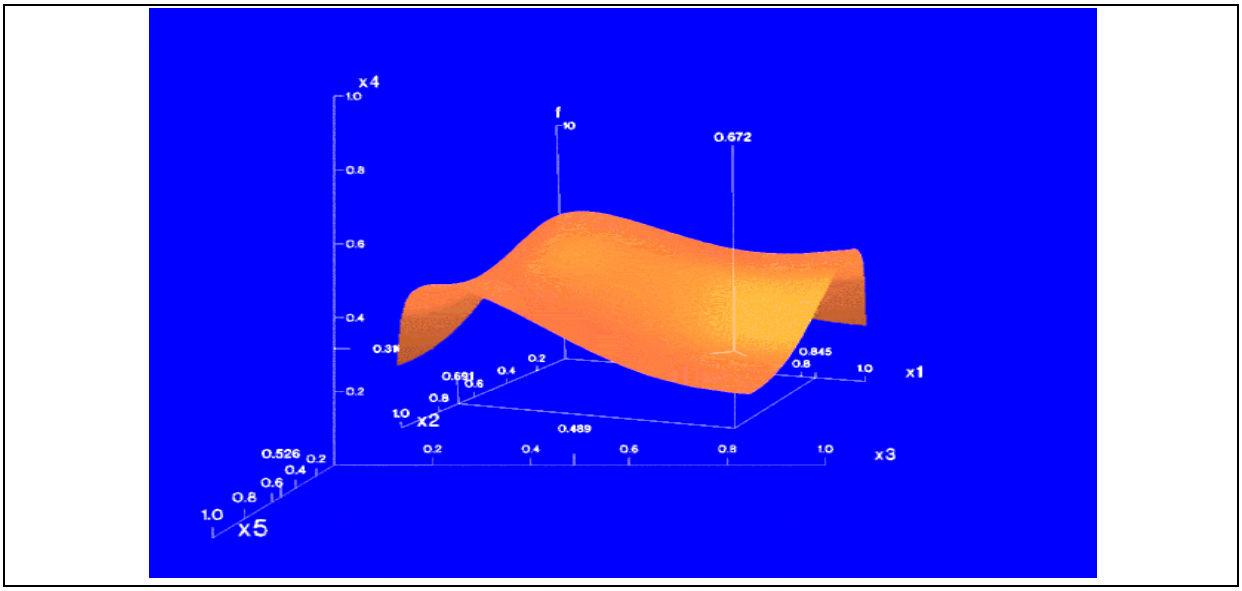


Εικόνα.3.4: Οπτικοποίηση με μήτρα διασποράς στην περίπτωση έξι (6) διαστάσεων (Πηγή. https://blog.biolab.si/2011/09/07/3d-visualizations-in-orange/).

3.3. Τεχνικές Ιεραρχικής Απεικόνισης

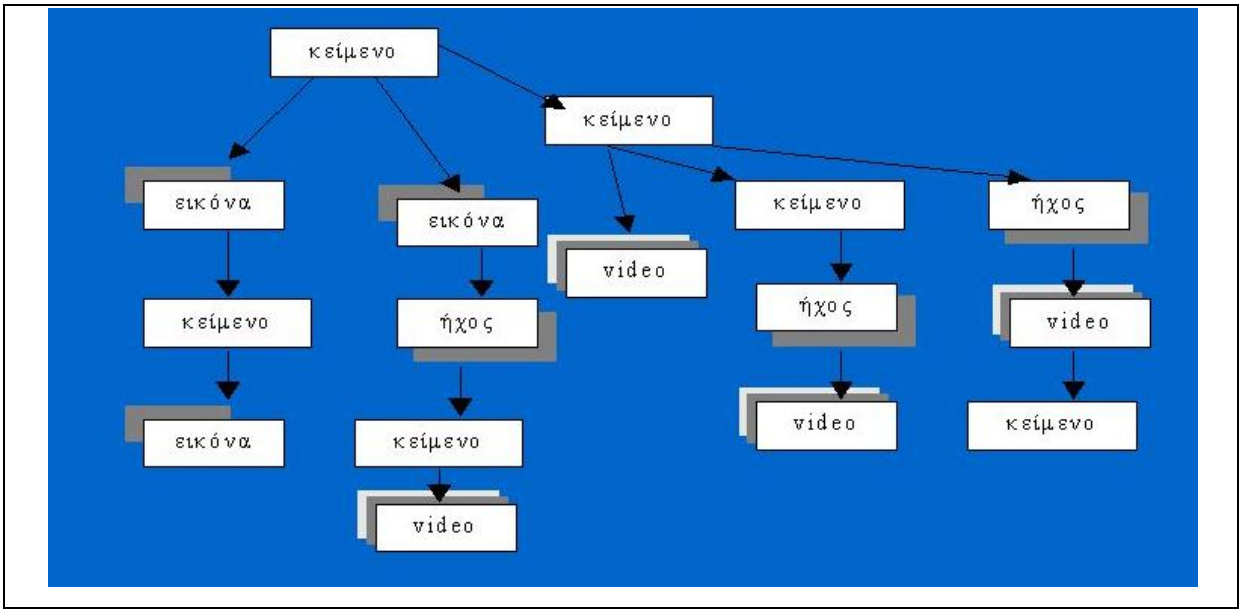
Οι τεχνικές απεικόνισης που συζητήθηκαν μέχρι τώρα επικεντρώνονται στην οπτικοποίηση πολλαπλών διαστάσεων ταυτόχρονα. Ωστόσο, για ένα μεγάλο σύνολο δεδομένων με πολλές διαστάσεις, αυτό δεν είναι πάντα εφικτό. Οι τεχνικές ιεραρχικής απεικόνισης χωρίζουν όλες τις διαστάσεις σε υποσύνολα. Το κάθε υποσύνολο του συνόλου των διαστάσεων αστικοποιείται διαδοχικά, με ένα ιεραρχικό τρόπο. Το γνωστό και ως **n-Vision**, είναι ένα αντιπροσωπευτικό παράδειγμα ιεραρχικής οπτικοποίησης η μορφή του οποίου φαίνεται και στην επόμενη εικόνα (Εικόνα 3.5). Ας υποτεθεί ότι πρέπει να απεικονιστεί ένα σύνολο δεδομένων 6-D, όπου οι διαστάσεις είναι {F, X1, ...,X5}. Επιθυμητό είναι η οπτικοποίηση να αναδεικνύει το πώς αλλάζει η διάσταση F σε σχέση με την καθεμία άλλη διάσταση.

Μπορούν πρώτα να καθοριστούν τις οι τιμές των διαστάσεων {X3, X4, X5} (ένα υποσύνολο του συνόλου των διαστάσεων), σε ορισμένο σημείο, π.χ. το σημείο (3, 4, 5). Μπορεί στη συνέχεια να απεικονιστεί το υποσύνολο {F, X1, X2} χρησιμοποιώντας ένα «χώρο» 3-D. Η αρχή των αξόνων για το εσωτερικό πεδίο {X3, X4, X5} βρίσκεται στο σημείο (3,4,5) ενώ το εξωτερικό πεδίο είναι ένας άλλος «χώρος» 3-D που απεικονίζεται χρησιμοποιώντας τις διαστάσεις {X3, X4, X5}. Ένας χρήστης μπορεί να αλληλεπιδρά με την αναπαράσταση αυτή, αλλάζοντας, στο εξωτερικό πεδίο, τη θέση της αρχής των αξόνων του εσωτερικού πεδίου. Ο χρήστης βλέπει τις προκύπτουσες αλλαγές του εσωτερικού πεδίου. Επιπλέον, ένας χρήστης μπορεί να ποικίλει τις διαστάσεις που χρησιμοποιούνται στο εσωτερικό πεδίο και στο εξωτερικό πεδίο. Για δεδομένα περισσότερων διαστάσεων, μπορούν να χρησιμοποιηθούν περισσότερα επίπεδα πεδίων, **[20], [22]**.



Εικόνα.3.5: n-Vision “Worlds within Worlds” (Πηγή. https://slidewiki.org/deck/1265-4/data-mining/slide/10481-2/1267-2:2;1277-2:5;1402-2:5;10481-2:4/view)

Ένα άλλο παράδειγμα μεθόδων ιεραρχικής απεικόνισης είναι οι δενδροειδείς απεικονίσεις, **[29]**, που εμφανίζουν ιεραρχική μορφή. Για παράδειγμα όπως βλέπουμε και στην συνέχεια (Εικόνα 3.6) σε μια αναζήτηση στο Google όλες οι ειδήσεις είναι οργανωμένες σε επτά κατηγορίες, όπου η κάθε μια εμφανίζεται ως ένα μεγάλο ορθογώνιο με μοναδικό χρώμα. Εντός κάθε κατηγορίας (δηλαδή κάθε ορθογώνιου στο ανώτερο επίπεδο) οι ειδήσεις διαχωρίζονται περαιτέρω σε μικρότερες υποκατηγορίες – ορθογώνια, και αυτό μπορεί να επαναλαμβάνεται σε πολλά ιεραρχικά χαμηλότερα επίπεδα, **[35]**.



Εικόνα.3.6: Δενδροειδής Δομή Πολυμέσων

(Πηγή. https://slideplayer.gr/slide/2853885)

3.4. Οπτικοποίηση Σύνθετων Δεδομένων

Αρχικά οι τεχνικές απεικόνισης δεδομένων αφορούσαν κυρίως αριθμητικά / ποσοτικά δεδομένα. Πρόσφατα, έχουν καταστεί διαθέσιμα περισσότερα ποιοτικά / μη ποσοτικά / μη αριθμητικά δεδομένα, όπως το κείμενο(text) και τα κοινωνικά δίκτυα. Η απεικόνιση και η ανάλυση τέτοιων δεδομένων προσελκύει μεγάλο ενδιαφέρον. Υπάρχουν πολλές νέες τεχνικές απεικόνισης αφιερωμένες σε αυτά τα είδη δεδομένων. Για παράδειγμα, πολλοί άνθρωποι στην ετικέτα του διαδικτύου επισημαίνουν διάφορα αντικείμενα όπως εικόνες, καταχωρήσεις ιστολογίου και κριτικές προϊόντων. Ένα «σύννεφο» ετικετών είναι μια απεικόνιση των στατιστικών των ετικετών που δημιουργούνται από τον συγκεκριμένο χρήστη. Συχνά, σε ένα σύννεφο ετικετών, οι ετικέτες παρατίθενται αλφαβητικά ή με μια σειρά που προτιμάται από τον χρήστη. Η σημαντικότητα μιας ετικέτας υποδεικνύεται από το μέγεθος ή το χρώμα της γραμματοσειράς όπως πχ ένα σύννεφο ετικετών για την απεικόνιση των δημοφιλών ετικετών που χρησιμοποιούνται σε μια τοποθεσία Web. Τα σύννεφα ετικετών χρησιμοποιούνται συχνά με δύο τρόπους. Πρώτον, σε ένα σύννεφο ετικετών για ένα μόνο στοιχείο, μπορεί χρησιμοποιηθεί το μέγεθος μιας ετικέτας για να

αναπαρασταθεί ο αριθμός των φορών που η ετικέτα εφαρμόζεται σε αυτό το στοιχείο από διαφορετικούς χρήστες.

Δεύτερον, κατά την απεικόνιση των στατιστικών ετικετών σε πολλά στοιχεία, μπορεί να χρησιμοποιηθεί το μέγεθος μιας ετικέτας για να αναπαρασταθεί ο αριθμός των στοιχείων στα οποία έχει εφαρμοστεί η ετικέτα, δηλαδή η δημοτικότητα (συχνότητα εμφάνισης) της ετικέτας.

Εκτός από τα πολύπλοκα δεδομένα, οι πολύπλοκες σχέσεις μεταξύ των καταχωρήσεων δεδομένων αποτελούν επίσης προκλήσεις για την οπτικοποίηση. Για παράδειγμα θα μπορούσαμε να είχαμε ένα γράφημα επιρροής μιας νόσου που να απεικονίζονται οι συσχετίσεις μεταξύ των ασθενειών. Οι κόμβοι στο γράφημα θα μπορούσαν να είναι ασθένειες και το μέγεθος κάθε κόμβου να είναι ανάλογο με την επικράτηση της αντίστοιχης νόσου. Δύο κόμβοι θα συνδέονταν με μία ακμή εάν οι αντίστοιχες ασθένειες είχαν ισχυρή συσχέτιση μεταξύ τους. Το πλάτος της ακμής μπορεί να είναι ανάλογο με τη δύναμη (ισχύ) του μοτίβου συσχέτισης των δύο συνδεόμενων ασθενειών, **[16], [13]**.

3.5. Εργαλεία Οπτικοποίησης

3.5.1. Gephi Graph Viz Platform

Το **Gephi [25]** είναι ένα πακέτο λογισμικού ανοιχτού κώδικα για ανάλυση και οπτικοποίηση δικτύων, το οποίο έχει γραφτεί σε γλώσσαJava στην πλατφόρμα NetBeans. Αρχικά αναπτύχθηκε από φοιτητές του Τεχνολογικού Πανεπιστημίου της Compiègne στη Γαλλία. Επιλέχθηκε για το Google Summer of Code τα έτη 2009, 2010, 2011, 2012 και 2013. Η τελευταία έκδοση του, v. 0.9.0, ξεκίνησε τον Δεκέμβριο του 2015 με ενημερώσεις τον Φεβρουάριο του 2016 (0.9.1) και τον Σεπτέμβριο του 2017 (0.9.2). Οι προηγούμενες εκδόσεις ήταν οι 0.6.0 (2008), 0.7.0 (2010) .0.8(2011), 0.8.1 (2012) και 0.8.2 (2013). Η Κοινοπραξία Gephi, που δημιουργήθηκε το 2010, είναι μια γαλλική μη κερδοσκοπική εταιρία που υποστηρίζει την ανάπτυξη μελλοντικών εκδόσεων του Gephi. Τα μέλη της περιλαμβάνουν τους φορείς SciencesPo, Linkfluence, WebAtlas και Quid. Το Gephi υποστηρίζεται επίσης από μια μεγάλη κοινότητα χρηστών, δομημένη σε μια ομάδα συζήτησης, **[24],** και ένα φόρουμ, **[24],** και που παράγει πολυάριθμες blog posts, έγγραφα και tutorials.

Το Gephi έχει χρησιμοποιηθεί σε πολλά ερευνητικά έργα στον ακαδημαϊκό χώρο, στη δημοσιογραφία και αλλού, για παράδειγμα στην οπτικοποίηση της παγκόσμιας συνδεσιμότητας του περιεχομένου των New York Times και στην εξέταση της κυκλοφορίας του δικτύου Twitter κατά τη διάρκεια κοινωνικών αναταραχών, μαζί με πιο παραδοσιακά θέματα δικτύωσης όπως οι ανθρωπιστικές επιστήμες (ιστορία, λογοτεχνία, πολιτικές επιστήμες κ.λπ.) – μια κοινότητα στην οποία εμπλέκονται πολλοί από τους υπεύθυνους για την ανάπτυξή του.

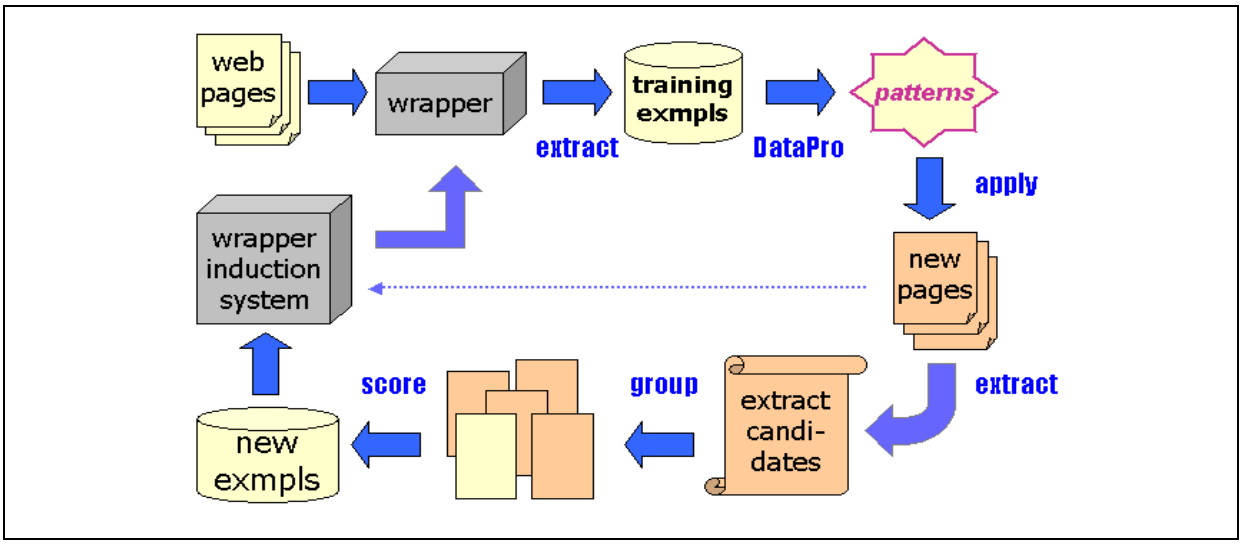
3.5.2. Datawrapper

Το **Datawrapper, [30],** είναι ένα εργαλείο απεικόνισης δεδομένων σχεδιασμένο για οργανισμούς ειδήσεων, για τη δημιουργία ενσωματωμένων απεικονίσεων μέσω μιας σχετικά απλής διαδικασίας. Ο χρήστης ακολουθεί μια διαδικασία τεσσάρων βημάτων, (μεταφόρτωση, έλεγχου και περιγραφή, οπτικοποίηση, δημοσίευση και ενσωμάτωση) για τη δημιουργία οπτικοποιήσεων μέσω του εργαλείου αυτού με την σειρά που βλέπουμε στην επόμενη εικόνα (Εικόνα 3.7). Το εργαλείο μπορεί να χρησιμοποιηθεί και για χαρτογράφηση. Το Datawrapper δημιουργεί δύο τύπους χαρτών: το Choropleth και το σημείο. Εάν δημιουργείται ένας χάρτη σημείων, η υπηρεσία επιτρέπει τη βασική γεωκωδικοποίηση, επιτρέποντας στον χρήστη να τοποθετεί ένα σημείο κάθε φορά πληκτρολογώντας μια διεύθυνση.

Οι προκύπτουσες απεικονίσεις είναι δυναμικές, καθώς το ποντίκι του Η/Υ μπορεί να αιωρείται πάνω από κάποιες απεικονίσεις για να πάρει περισσότερες πληροφορίες. Η υπηρεσία είναι ελεύθερη να χρησιμοποιηθεί έως ένα ορισμένο επίπεδο προβολών, μετά το οποίο ο χρήστης θα πρέπει να εγγραφεί. Το Datawrapper δημιουργεί διαδραστικές οπτικοποιήσεις που μπορούν να ενσωματωθούν σε έναν ιστότοπο. Η δύναμη του εργαλείου είναι η ευκολία χρήσης του και η ταχύτητα με την οποία μπορεί να δημιουργήσει μια απεικόνιση, αλλά αυτό συμβαίνει με κόστος την αυξημένη πολυπλοκότητα πέρα από το πεδίο εφαρμογής του εργαλείου.

Η διεπαφή είναι απλούστερη από τη χρήση των φύλλων Google, αλλά παρέχει πολύ λίγη υποστήριξη για την επεξεργασία των δεδομένων και προϋποθέτει ότι έχει ήδη γίνει η διαδικασία προ-επεξεργασίας των δεδομένων πριν χρησιμοποιηθεί το εργαλείο.

Το **Tableau Public, [27],** είναι ένα πιο ισχυρό εργαλείο και δημιουργεί επίσης διαδραστικές απεικονίσεις που μπορούν να ενσωματωθούν, αλλά η διεπαφή είναι πιο περίπλοκη και είναι ένα εργαλείο επιφάνειας εργασίας. Ως εργαλείο χαρτογράφησης, το Datawrapper παρέχει καθαρά, βασικά χαρακτηριστικά. Σε αντίθεση με κάποια εργαλεία χαρτογράφησης, μπορεί να δημιουργήσει μια choroplethual απεικόνιση για διάφορα μέρη. Δεν μπορεί να δημιουργήσει έναν χάρτη σημείων από ένα προϋπάρχον σύνολο δεδομένων όπως τα Carto ή Tableau Public.

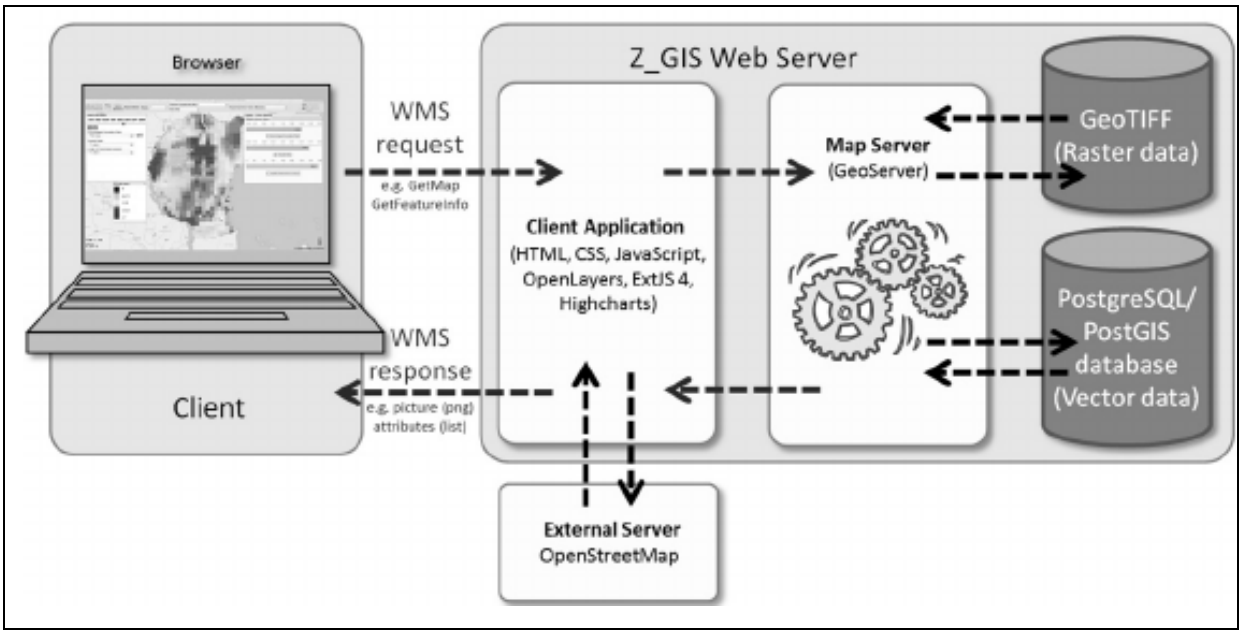


Εικόνα.3.7: Διαδικασία δημιουργίας Ενσωματωμένων Απεικονίσεων
(Πηγή. https://www.isi.edu/integration/Mercury/)

3.5.3. Highcharts

Το **Highcharts, [31],** είναι ένα προϊόν που δημιουργήθηκε από την εταιρεία Highsoft που εδρεύει στη Νορβηγία. Το Highcharts κυκλοφόρησε το 2009 και είναι μια βιβλιοθήκη γραφικών που γράφτηκε με καθαρή γλώσσα JavaScript και αλληλεπιδρά με το διαδίκτυο όπως φαίνεται στην συνέχεια (Εικόνα 3.8). Το προϊόν αναπτύσσεται στο Vik της Νορβηγίας και παρουσιάζεται τακτικά στα εθνικά μέσα ενημέρωσης, όπως οι Finansavisen και Dagsrevyen. Ο Torstein Hønsi είναι ο κύριος δημιουργός του προϊόντος. Το προϊόν παρουσιάστηκε για πρώτη φορά και αναθεωρήθηκε το 2006. Σε μια συνέντευξή του, ο Finansavisen μίλησε για την ανάγκη ενός λογισμικού για τη δημιουργία γραφημάτων, επιτρέποντας στους χρήστες να αναρτούν γραφήματα απευθείας σε έναν ιστότοπο. Σε αντίθεση με πολλά προϊόντα λογισμικού, δεν αναπτύσσεται σε γνωστή τεχνολογική

τοποθεσία, όπως η Silicon Valley. Η εταιρεία εδρεύει στη μικρή νορβηγική πόλη Vik στη Νορβηγία, εντούτοις το προϊόν χρησιμοποιείται παγκοσμίως. Αναφέρθηκε σε μια συνέντευξη το 2014 ότι η πελατειακή του βάση είναι κυρίως από το εξωτερικό, με το 97% των εσόδων της εταιρείας να προέρχεται από πελάτες εκτός της Νορβηγίας. Τα προϊόντα Highcharts εμφανίστηκαν στη νορβηγική τηλεόραση στα τέλη του 2012, όταν η μητρική εταιρεία εμφανίστηκε σε μια εθνική τηλεοπτική εκπομπή. Ήταν ένα τμήμα στη μέση μιας επίδειξης 30 λεπτών, από τη Dagsrevyen. Το Highcharts είναι δωρεάν για μη εμπορική και προσωπική χρήση, με την υποχρέωση να αγοραστούν άδειες μόνο για εμπορικές εφαρμογές.



Εικόνα.3.8: Αλληλεπίδραση του Highcharts με το διαδίκτυο

(Πηγή. https://www.researchgate.net/figure/The-WebGIS-architecture-with-Client-Server-Data-structure_fig1_279537223).

3.5.4. Plotly

Το **Plotly, [26],** (https://plot.ly/) η δομή του οποίου φαίνεται στην επόμενη εικόνα (Εικόνα 3.9), το δημιούργησε μια εταιρεία τεχνικών υπολογιστών που εδρεύει στο Μόντρεαλ του Κεμπέκ στον Καναδά. Η εταιρία αναπτύσσει ηλεκτρονικά στοιχεία ανάλυσης και εργαλεία οπτικοποίησης. Το Plotly παρέχει online εργαλεία γραφικών, αναλυτικών και στατιστικών στοιχείων για άτομα καθώς και επιστημονικές βιβλιοθήκες γραφικών για τις

γλώσσες προγραμματισμού Python, R, MATLAB, Perl, Julia, Arduino κτλ. Το Plotly ιδρύθηκε από τον Alex Johnson, τον Jack Parmer, τον Chris Parmer και τον Matthew Sundquist. Τα επιστημονικά υπόβαθρα των ιδρυτών του εντοπίζονται στην επιστήμη, την ενέργεια, την ανάλυση δεδομένων και την απεικόνιση.

Οι πρώτοι που δημιούργησαν το Plotly ήταν οι Christophe Viau, ένας καναδός μηχανικός λογισμικού και ο Ben Postlethwaite, ένας καναδός γεωφυσικός. Το Plotly ονομάστηκε μία από τις Top 20 Hottest Innovative Companies στον Καναδά από την Καναδική Ανταλλαγή Καινοτομίας. Το Plotly πρώτο-παρουσιάστηκε στην "σειρά εκκίνησης" στο συνέδριο PyCon 2013 και κατασκευάστηκε χρησιμοποιώντας τη γλώσσα Python και το πλαίσιο Django, με ένα front-end που χρησιμοποιεί τη JavaScript και τη βιβλιοθήκη οπτικοποίησης D3.js, καθώς και τις γλώσσες HTML και CSS.

Εικόνα.3.9: Εργαλεία γραφικών, Αναλυτικών & Στατιστικών στοιχείων.

(Πηγή. https://plot.ly/products/on-premise/)

4.1. Χρησιμότητα της Αναλυτικής Δεδομένων (Data Analytics)

Τα δεδομένα και οι αναλύσεις επιτρέπουν στους ερευνητές να λαμβάνουν τεκμηριωμένες αποφάσεις και να μην καταφεύγουν σε «μαντείες». Τα δεδομένα «είναι αυτά που είναι» (αντικειμενικότητα): ακόμη και αν μπορεί εύκολα να καταστρατηγηθούν, θα δίνουν πάντα την αλήθεια. Η ανάλυση δεδομένων παρέχει αντικειμενικές απαντήσεις που μπορούν να οδηγήσουν στη λύση ενός προβλήματος ή την επιτυχία ενός εγχειρήματος. Το πρόσθετο πλεονέκτημα που προσφέρουν αφορά ακριβώς τους αναλυτές πληροφορικής. Οι επιχειρήσεις πρέπει να κάνουν συμβιβασμούς. Οι αεροπορικές εταιρείες θα μπορούσαν να ανταλλάξουν απόδοση για φορτίο, ή το αντίστροφο. Τα ταξιδιωτικά γραφεία πρέπει να ξοδεύουν τον διαφημιστικό προϋπολογισμό τους με το μέγιστο δυνατό αποτέλεσμα. Τα δεδομένα και τα αναλυτικά στοιχεία μπορούν να επηρεάσουν πραγματικά τις αποφάσεις που λαμβάνει μια επιχείρηση και το αποτέλεσμά τους, **[2]**.

Εκτός από τον επιχειρηματικό αντίκτυπο, υπάρχουν και άλλοι λόγοι για την ανάλυση του καινούργιου αυτού τομέα. Τα εργαλεία για το χειρισμό μεγάλων ποσοτήτων δεδομένων και τη διεξαγωγή αναλυτικών στοιχείων πιθανότατα προχωρούν πιο γρήγορα από οποιοδήποτε άλλο πεδίο τεχνολογίας σήμερα. Με καλύτερα εργαλεία, είναι ουσιαστικότερη και βαθύτερη η κατανόηση του τρόπου με τον οποίο πρέπει να γίνονται οι αναλύσεις και πώς πρέπει να χρησιμοποιούνται οι πληροφορίες που προκύπτουν. Ας υποτεθεί ότι σε ένα αεροδρόμιο κάποιος παρατηρεί τα αεροσκάφη να έρχονται και να φεύγουν, να φεύγουν από τις πύλες, να επιβιβάζονται επιβάτες και να απογειώνονται ξανά, κτλ. Όποια ερωτήματα κι αν έρθουν στο νου του παρατηρητή, είναι πιθανότατο ότι η ανάλυση των δεδομένων που κυκλοφορούν στα πληροφορικό συστήματα του αεροδρόμιου μπορεί να τα απαντήσει γρήγορα και με ακρίβεια.

Τι είναι η επιστήμη των δεδομένων (data science) ή με νεώτερους όρους η αναλυτική των δεδομένων (data analytics); Σε ένα τόσο νέο πεδίο, η εξεύρεση κοινής άποψης γύρω από έναν ενιαίο ορισμό είναι δύσκολη. Αλλά ένας ορισμός που θα μπορούσε να προσεγγίσει την πραγματικότητα θα μπορούσε να ήταν ότι Αναλυτική Δεδομένων είναι η

διαδικασία δημιουργίας αξίας από τα δεδομένα. Όχι μόνο η κατανόηση των δεδομένων ή η πραγματοποίηση προβλέψεων με βάση τα δεδομένα αλλά το ίδιο το γεγονός ότι υπάρχουν τόσα πολλά στοιχεία αποθηκευμένα και διαθέσιμα, που ακόμη και κάτι τόσο απλό όσο ο υπολογισμός των τελικών ναύλων π.χ. απαιτεί εξειδικευμένη τεχνολογία, κάνει τον τομέα των Data Analytics έναν ανερχόμενο δυναμικό τομέα, **[2], [7]**.

Μπορεί να εφαρμοστεί σε πολλές διαφορετικές επιστημονικές ή μη περιοχές. Στα δεδομένα της βιομηχανίας ταξιδιών (εισιτήρια, χρονοδιαγράμματα, κρατήσεις, αναζητήσεις), σε οικονομικά και χρηματιστηριακά δεδομένα όπως οι συναλλαγματικές ισοτιμίες, στα δεδομένα μετεωρολογίας και πρόβλεψης καιρού και κυρίως στα εκπαιδευτικά δεδομένα που είναι και το επίκεντρο ενδιαφέροντος της παρούσας εργασίας (όπως επιδόσεις στα μαθήματα είτε ανά κεφάλαιο είτε συνολικά, ενδιαφέροντα των φοιτητών σε κάποιο μάθημα ή σε κάποια ενότητα ξεχωριστά, κτλ). Το σίγουρο είναι ότι η επιστήμη των δεδομένων ως ευρύτερος υποτομέας της επιστήμης της πληροφορικής αποτελεί ένα πολύ δυνατό εργαλείο για την ανάλυση δεδομένα με σκοπό πάντα την βελτίωση ενός συστήματος, φυσικού ή κοινωνικού, π.χ. μίας κοινότητας μάθησης στα εκπαιδευτικά πλαίσια.

4.2. Ταξινόμηση (Classification)

Η ταξινόμηση (classification) είναι μια μορφή ανάλυσης δεδομένων που εξάγει μοντέλα που περιγράφουν σημαντικές κατηγορίες δεδομένων. Τέτοια μοντέλα, που ονομάζονται ταξινομητές (classifiers), εντάσσουν τα επιμέρους δεδομένα εισόδου τους σε συγκεκριμένες κατηγορίες (διακριτές, μη διατεταγμένες) αποδίδοντας στο καθένα μία ετικέτα κλάσης. Για παράδειγμα, μπορεί να δημιουργηθεί ένα μοντέλο ταξινόμησης για να κατηγοριοποιηθούν οι αιτήσεις τραπεζικών δανείων. Μια τέτοια ανάλυση μπορεί να βοηθήσει στην καλύτερη κατανόηση των δεδομένων που συλλέγει και τηρεί η τράπεζα, **[7], [14]**.

Πολλές μέθοδοι ταξινόμησης έχουν προταθεί από τους ερευνητές στη μηχανική μάθηση και στην αναγνώριση προτύπων με στατιστικά στοιχεία. Οι περισσότεροι αλγόριθμοι είναι δυνατόν να «τρέξουν» κεντρικά σε έναν υπολογιστή, τυπικά υποθέτοντας ένα μικρό όγκο δεδομένων. Πρόσφατη έρευνα για το Data Mining έχει βασιστεί στην

ανάπτυξη κλιμακούμενων τεχνικών ταξινόμησης και πρόβλεψης ικανών να χειρίζονται ιδιαίτερα μεγάλες ποσότητες δεδομένων (big data).

Η ταξινόμηση υπηρετεί πολλές άλλες σημαντικές πρακτικές εφαρμογές, όπως ανίχνευση απάτης, marketing των στόχων, πρόβλεψη απόδοσης, κατασκευή συστημάτων και ιατρική διάγνωση. Επειδή η ανάλυση των βασικών τεχνικών ταξινόμησης είναι εκτός του πλαισίου της παρούσας εργασίας, στη συνέχεια εξετάζονται εντελώς επιγραμματικά οι βασικές τεχνικές για την ταξινόμηση δεδομένων:

- ✓ Ταξινομητές δέντρων αποφάσεων,
- ✓ Ταξινομητές τύπου Bayes,
- ✓ Ταξινομητές τύπου κανόνων συσχέτισης

και τέλος θα γίνει περιγραφή τον τρόπου αξιολόγησης και σύγκρισης διαφορετικών ταξινομητών.

Υπάρχουν διάφορα μέτρα της επιδιωκόμενης ακρίβειας καθώς και τεχνικές για να επιτυγχάνονται αξιόπιστες εκτιμήσεις της ακρίβειας. Υπάρχουν και μέθοδοι για την αύξηση της ακρίβειας ενός ταξινομητή περιλαμβανομένων περιπτώσεων για το πότε το σύνολο δεδομένων είναι ισορροπημένο ως προς τις κλάσεις που περιλαμβάνει (καθώς ενδέχεται η κύρια κατηγορία ενδιαφέροντος να είναι σπάνια). Ένας διευθυντής μάρκετινγκ π.χ. μιας εταιρείας χρειάζεται ανάλυση δεδομένων για να μάθει αν ένας πελάτης με ένα δεδομένο προφίλ θα αγοράσει έναν νέο υπολογιστή. Ένας ιατρικός ερευνητής θέλει να αναλύσει τα στοιχεία του καρκίνου του μαστού για να ορίσει συγκεκριμένες θεραπείες που θα πρέπει να λαμβάνει ένας ασθενής. Σε κάθε ένα από αυτά τα παραδείγματα ανάλυσης δεδομένων, το βασικό πρώτο βήμα είναι η ταξινόμηση, όπου ένα μοντέλο ή ταξινομητής κατασκευάζεται για να προβλέψει την κατηγορία όπου ανήκει κάθε ένα συγκεκριμένο δεδομένο εισόδου. Οι έξοδοι του ταξινομητή μπορεί να είναι δυαδικού τύπου, όπως "Ναι" ή "Όχι" για τα δεδομένα στο παράδειγμα του μάρκετινγκ, ή τύπου «1 από Ν», όπως "θεραπεία Α", "θεραπεία Β" ή "θεραπεία C" για το παράδειγμα με τα ιατρικά δεδομένα κτλ., **[2], [10], [12]**.

Αυτές οι κατηγορίες μπορούν να εκπροσωπούνται από διακριτές τιμές, όπου η σειρά (διάταξη) μεταξύ των αξιών δεν έχει νόημα. Για παράδειγμα, ο στόχος της ανάλυσης δεδομένων μπορεί είναι ένα πρόβλημα αριθμητικής πρόβλεψης, όπου το κατασκευασμένο μοντέλο προβλέπει τις (μελλοντικές) τιμές μιας συνάρτησης που λαμβάνει τιμές σε ένα συνεχές σύνολο.

Ανάλυση παλινδρόμησης (Regression Analysis) είναι μια στατιστική μέθοδος που χρησιμοποιείται συχνότερα για την αριθμητική πρόβλεψη, εξ ου και οι δύο όροι τείνουν να χρησιμοποιούνται σχεδόν συνώνυμα. Η ταξινόμηση και η αριθμητική πρόβλεψη είναι οι δύο κύριοι τύποι πρόβλεψης που χειρίζονται συνήθως κοινούς τύπους προβλημάτων.

Η ταξινόμηση δεδομένων είναι μια διαδικασία δύο σταδίων, η οποία αποτελείται από ένα κατασκευαστή μοντέλου (κατασκευάζεται ένα μοντέλο ταξινόμησης, δηλαδή ένας ταξινομητής, συνήθως σε μορφή αλγορίθμου κωδικοποιημένου σε κάποια γλώσσα λογισμικού) και στη συνέχεια το κυρίως βήμα της ταξινόμησης (όπου το μοντέλο χρησιμοποιείται για την πρόβλεψη ετικετών κλάσης για νέα, άγνωστα δεδομένα εισόδου).

Στο πρώτο βήμα, δημιουργείται ένας ταξινομητής που στηρίζεται σε ένα προκαθορισμένο σύνολο κατηγοριών δεδομένων ή εννοιών. Αυτό είναι το βήμα (ή η φάση) εκπαίδευσης ή μάθησης ή κατάρτισης, κατά το οποίο ένας αλγόριθμος ταξινόμησης «χτίζει» τον ταξινομητή αναλύοντας ή "μαθαίνοντας" τα πρότυπα που θα πρέπει μελλοντικά να ταξινομεί, με βάση ένα σύνολο παραδειγμάτων εκπαίδευσης (training set) που αποτελείται από ένα σύνολο δεδομένων εισόδου, το καθένα συνοδευόμενο (συνδεδεμένο) με τη σχετική «ετικέτα» (label) της ορθής κλάσης στην οποία ανήκει. Μια πλειάδα, έστω X, αντιπροσωπεύεται από ένα πίνακα **n**-διαστάσεων που απεικονίζει τις μετρήσεις που έγιναν στην πλειάδα από **n** χαρακτηριστικά από τη βάση δεδομένων, αντίστοιχα, π.χ. {A1, A2, …., An}. Κάθε πλειάδα X ανήκει σε μια προκαθορισμένη κλάση (1 από N), που προσδιορίζεται από ένα άλλο χαρακτηριστικό που ονομάζεται χαρακτηριστικό ετικέτας.

Στο δεύτερο βήμα, το μοντέλο-ταξινομητής που προέκυψε από την πρώτη φάση χρησιμοποιείται για ταξινόμηση άγνωστων δεδομένων (πλειάδων) εισόδου, που δεν συμμετείχαν στο training set και για οι οποίες δεν συνοδεύονται από την ορθή ετικέτα κλάσης. Στο βαθμό που η φάση εκπαίδευσης ήταν επιτυχημένη, ο ταξινομητής θα μπορέσει να «γενικεύσει» (generalization property) τα συμπεράσματά του, ώστε να λειτουργεί σωστά με μεγάλη πιθανότητα και για το ευρύτερο σύνολο των άγνωστων εισόδων, **[2], [7], [10], [16]**.

Σημαντικό είναι μετά το πέρας της πρώτης φάσης, να εκτιμάται η προβλεπτική ακρίβεια του ταξινομητή. Αν για το σκοπό αυτό χρησιμοποιηθεί το ίδιο το training set, αυτή η εκτίμηση είναι πιθανό να είναι αισιόδοξη, διότι ο ταξινομητής έχει ήδη εκπαιδευθεί να ταξινομεί σωστά τα δεδομένα αυτά. Επιπλέον κίνδυνος υπεραισιόδοξης εκτίμησης της

ακρίβειάς του οφείλεται στην τάση του ταξινομητή να υπερκεράσει τα δεδομένα (αυτό μπορεί να συμβεί αν, κατά τη φάση εκπαίδευσης, δώσει βαρύτητα και ενσωματώσει ορισμένες ιδιαίτερες ανωμαλίες των δεδομένων του training set που δεν υπάρχουν στο γενικό σύνολο των άγνωστων δεδομένων που θα αντιμετωπίσει στη συνέχεια). Οπότε για την εκτίμηση της ακρίβειας χρησιμοποιείται συνήθως ένα υποσύνολο των διαθέσιμων δεδομένων της βάσης που δεν είχε αξιοποιηθεί κατά τη φάση εκπαίδευσης, ακριβώς για να είναι άγνωστο στον ταξινομητή και να μπορέσει να ελεγχθεί σωστά πάνω σε αυτό η ακρίβειά του (το ποσοστό ορθών κατηγοριοποιήσεων % πάνω στο άγνωστο σύνολο εισόδων).

4.3. Κανόνες Συσχέτισης (Association Rules)

Ένας κανόνας συσχέτισης (Association Rule) αποτελείται από μια πρόταση αριστερής πλευράς – (Left-hand Side – LHS) και μια πρόταση δεξιάς πλευράς – (Right-hand Side – RHS). Και οι δύο πλευρές αποτελούνται από δηλώσεις δυαδικού (Boolean) τύπου, δηλαδή προτάσεις που είτε ισχύουν (λογικό «1») είτε δεν ισχύουν (λογικό «0»). Στον προγραμματισμό Η/Υ η μία τέτοια δήλωση χαρακτηρίζεται ως δήλωση τύπου «Αλήθεια / Ψέμα – True / False ή T/F). Ο κανόνας υπονοεί ότι αν η πρόταση στην αριστερή πλευρά είναι αλήθεια, τότε η πρόταση στη δεξιά πλευρά είναι επίσης αλήθεια. Συχνά εισάγεται ένας κανόνας πιθανότητας, ώστε η δεξιά πλευρά να είναι αληθής με την πιθανότητα p, δεδομένου ότι η αριστερή πλευρά είναι αληθής με πιθανότητα 1.

Ένας τυπικός ορισμός του κανόνα συσχέτισης δίνεται παρακάτω. Ένας κανόνας συσχέτισης είναι ένας κανόνας με τη μορφή

$$X => Y$$

όπου τα X και Y είναι πρότυπα ή σύνολα αντικειμένων. Καθώς ο αριθμός των παραγόμενων συσχετίσεων μπορεί να είναι τεράστιος, οι συσχετίσεις έχουν νόημα με δύο μέτρα πιθανότητας, που ονομάζονται «υποστήριξη» και «εμπιστοσύνη» και εισάγονται για να απορρίπτουν τις λιγότερο συχνές συσχετίσεις μέσα στη βάση δεδομένων. Η υποστήριξη είναι η πιθανότητα να βρεθούν X και Y στην ίδια ομάδα, ενώ η εμπιστοσύνη είναι η υπό όρους πιθανότητα να βρεθεί μια ομάδα Y που να έχει το X. Οι τυπικοί ορισμοί της υποστήριξης και της εμπιστοσύνης δίνονται παρακάτω.

Λαμβάνοντας ένα σύνολο στοιχείων (πλειάδων) X, η συχνότητα fr(X) είναι ο αριθμός των περιπτώσεων εντός των δεδομένα που ικανοποιούν το X (ή για τα οποία η πρόταση X ισχύει, αληθεύει).

- Η υποστήριξη είναι η συχνότητα fr(X Λ Y).
- Η εμπιστοσύνη είναι το κλάσμα των πλειάδων που ικανοποιούν την πρόταση Y μεταξύ εκείνων των πλειάδων που ικανοποιούν την πρόταση X.

Οι κανόνες συσχέτισης συγκαταλέγονται μεταξύ των δημοφιλέστερων λύσεων για τα τοπικά πρότυπα στο Data Mining. Ο αλγόριθμος **Apriori** είναι ένας από τους πρώτους για εύρεση κανόνων συσχέτισης. Είναι ένας αλγόριθμος επιρροής για συχνά εμφανιζόμενες ομάδες αντικειμένων εξόρυξης ανάμεσα στους κανόνες συσχέτισης τύπου Boolean. Απαιτεί έναν αριθμό προσβάσεων στη βάση δεδομένων. Κατά τη διάρκεια της πρόσβασης k, ο αλγόριθμος βρίσκει το σύνολο των πλέον συχνά εμφανιζόμενων αντικειμένων Lk μήκους k που πληρούν την ελάχιστη απαίτηση ως προς την τιμή της υποστήριξης. Ο αλγόριθμος τερματίζεται όταν το Lk είναι κενό, **[2], [7], [10], [16], [35].**

4.4. Πρόβλεψη (Prediction)

Η ικανότητα πρόβλεψης της επίδοσης ενός μαθητή είναι πολύ σημαντικό στοιχείο σε εκπαιδευτικά περιβάλλοντα. Η επίδοση ενός ακαδημαϊκού φοιτητή βασίζεται σε ποικίλους παράγοντες όπως προσωπικούς, κοινωνικούς, ψυχολογικούς και άλλες περιβαλλοντικές μεταβλητές. Πολύ ελπιδοφόρο εργαλείο για την επίτευξη του στόχου της αποτίμησης ή της πρόβλεψης της ακαδημαϊκής επίδοσης είναι η χρήση του Data Mining. Όπως αναλύθηκε στα προηγούμενα, οι τεχνικές εξόρυξης δεδομένων χρησιμοποιούνται σε μεγάλες ποσότητες δεδομένων για να ανακαλύψουμε κρυμμένα μοτίβα και σχέσεις που θα έχουν χρησιμότητα στη λήψη αποφάσεων. Στην πραγματικότητα, μία από τις πιο χρήσιμες τεχνικές εξόρυξης δεδομένων στο πλαίσιο της ηλεκτρονικής μάθησης είναι η ταξινόμηση όμως εξίσου χρήσιμα είναι και τα προγνωστικά μοντέλα που έχουν τον ειδικό στόχο να επιτρέψουν την πρόβλεψη πάνω στις τιμές μεταβλητών ενδιαφέροντος όταν δίδονται γνωστές τιμές άλλων μεταβλητών.

Η προγνωστική μοντελοποίηση συχνά αναφέρεται ως εποπτευόμενη μάθηση (supervised learning) επειδή οι κλάσεις προσδιορίζονται πριν εξεταστούν τα δεδομένα. Τα πρότυπα πρόβλεψης περιλαμβάνουν όλα τα προσωπικά, κοινωνικά, ψυχολογικά δεδομένα

των μαθητών ή των φοιτητών καθώς και άλλες περιβαλλοντικές μεταβλητές που απαιτούνται για την αποτελεσματική πρόβλεψη της επίδοσής τους. Η πρόβλεψη της επίδοσης των μαθητών με υψηλή ακρίβεια είναι χρήσιμη για τον εντοπισμό των μαθητών με χαμηλό επίπεδο και την έγκαιρη ειδοποίηση και ενίσχυσή τους και στη συνέχεια για την παρακολούθηση των ακαδημαϊκών επιτευγμάτων.

Στο πλαίσιο αυτό, οι στόχοι της σχετικής έρευνες έχουν διαμορφωθεί έτσι ώστε να βοηθήσουν το ακαδημαϊκό φοιτητικό κοινό χαμηλών επιδόσεων στην τριτοβάθμια εκπαίδευση με τα εξής βήματα:

- ✓ Δημιουργία μιας πηγής δεδομένων προγνωστικών μεταβλητών
- ✓ Προσδιορισμός διαφορετικών παραγόντων, οι οποίοι επηρεάζουν τη μαθησιακή συμπεριφορά και την απόδοση κατά την ακαδημαϊκή σταδιοδρομία
- ✓ Κατασκευή μοντέλου πρόβλεψης με χρήση δεδομένων ταξινόμησης και τεχνικών εξόρυξης δεδομένων από τη βάση και
- ✓ Επικύρωση του μοντέλου για την τριτοβάθμια εκπαίδευση με βάση πραγματικά στοιχεία από φοιτητές που φοιτούν σε πανεπιστήμια ή άλλα ακαδημαϊκά ιδρύματα **[7], [11], [16]**.

4.5. Ομαδοποίηση ή Ανάλυση Συμπλέγματος (Clustering)

Ανάλυση συμπλέγματος ή απλά ομαδοποίηση (clustering) είναι η διαδικασία διαίρεσης ενός συνόλου αντικειμένων (δεδομένων) σε υποσύνολα. Κάθε υποσύνολο συγκροτεί ένα σύμπλεγμα ή ομάδα (cluster) και η ομαδοποίηση γίνεται έτσι ώστε αντικείμενα στο ίδιο σύμπλεγμα να είναι παρόμοια μεταξύ τους, αλλά αντικείμενα σε διαφορετικά συμπλέγματα να διαφέρουν ριζικά μεταξύ τους. Το σύνολο των ομάδων που προκύπτει από μια τέτοια διαδικασία τοποθέτησης των αντικειμένων σε ομάδες αναφέρεται ως ομαδοποίηση. Στο πλαίσιο αυτό διαφορετικές μέθοδοι ομαδοποίησης μπορεί να καταλήξουν σε διαφορετικές ομαδοποιήσεις για το ίδιο σύνολο δεδομένων, αν χρησιμοποιούν διαφορετικά κριτήρια ομοιότητας/διαφοράς. Η κατανομή δεν γίνεται από τον άνθρωπο, αλλά από τον αλγόριθμο ομαδοποίησης, **[2], [7]**.

Ως εκ τούτου η ομαδοποίηση είναι χρήσιμη διότι μπορεί να οδηγήσει στην ανακάλυψη προηγουμένως άγνωστων ομοιοτήτων εντός των δεδομένων. Η ομαδοποίηση ή ανάλυση συμπλέγματος έχει χρησιμοποιηθεί ευρέως σε πολλές εφαρμογές, όπως η

επιχειρηματική ευφυΐα, η αναγνώριση προτύπων εικόνας, η αναζήτηση στο Web, η βιολογία και η ασφάλεια. Στις επιχειρήσεις, η συλλογή πληροφοριών μπορεί να χρησιμοποιηθεί για την οργάνωση μεγάλου αριθμού πελατών σε ομάδες, όπου οι πελάτες μίας ομάδας έχουν ισχυρά παρόμοια χαρακτηριστικά. Αυτό διευκολύνει την ανάπτυξη επιχειρηματικών στρατηγικών για βελτιωμένη διαχείριση πελατειακών σχέσεων.

Στην αναγνώριση εικόνων, η ομαδοποίηση μπορεί να χρησιμοποιηθεί για να ανακαλύψει ομάδες ή "υποκατηγορίες" στα συστήματα αναγνώρισης χειρόγραφων χαρακτήρων.

Η ομαδοποίηση έχει επίσης βρει πολλές εφαρμογές στην αναζήτηση στο Web. Για παράδειγμα, μια λέξη-κλειδί στη αναζήτηση μπορεί συχνά να επιστρέφει έναν πολύ μεγάλο αριθμό ιστοσελίδων σχετικών με την αναζήτηση. Η ομαδοποίηση μπορεί να χρησιμοποιηθεί για την οργάνωση των αποτελεσμάτων σε ομάδες, ώστε αυτά να παρουσιαστούν με έναν συνοπτικό και εύκολα αντιληπτό τρόπο στον χρήστη. Επιπλέον, έχουν αναπτυχθεί τεχνικές ομαδοποίησης για τη συγκέντρωση εγγράφων πάνω σε συγκεκριμένα θέματα, που χρησιμοποιούνται συνήθως στην ανάκτηση πληροφοριών.

Στο πλαίσιο του Data Mining, το Clustering μπορεί να χρησιμοποιηθεί ως αυτόνομο εργαλείο για την κατανόηση της κατανομής των δεδομένων την εξαγωγή των κοινών χαρακτηριστικών κάθε ομάδας και την επικέντρωση σε ένα συγκεκριμένο σύνολο ομάδων για περαιτέρω ανάλυση. Εναλλακτικά, μπορεί να χρησιμεύσει ως στάδιο προ-επεξεργασίας για άλλους αλγόριθμους, όπως ο χαρακτηρισμός ή η εξαγωγή χαρακτηριστικών, η επιλογή και η ταξινόμηση, που στη συνέχεια θα λειτουργήσουν πάνω στα ανιχνευόμενα συμπλέγματα.

Επειδή ένα σύμπλεγμα είναι μια συλλογή αντικειμένων ή δεδομένων που είναι παρόμοια μεταξύ τους ενώ διαφέρουν από τα αντικείμενα σε άλλα συμπλέγματα, μπορεί ένα σύνολο αντικειμένων ή δεδομένων να αντιμετωπίζεται ως μια «σιωπηρή» τάξη ή κατηγορία. Με αυτή την έννοια, η ομαδοποίηση αποκαλείται μερικές φορές και αυτόματη ταξινόμηση ή ταξινόμηση χωρίς επίβλεψη (unsupervised classification).

Χρησιμοποιείται επίσης ο όρος «κατάτμηση δεδομένων» (data segmentation) σε ορισμένες εφαρμογές, λόγω της ομαδοποίησης που χωρίζει τα μεγάλα σύνολα δεδομένων σε ομάδες ανάλογα με την ομοιότητά τους. Ομαδοποίηση μπορεί επίσης να χρησιμοποιηθεί για ανίχνευση των ακραίων τιμών (outliers), όπου οι ακραίες τιμές (τιμές που είναι "μακριά" από οποιεσδήποτε ομάδα) μπορεί να είναι πιο ενδιαφέρουσες από τις

συνήθεις περιπτώσεις. Εφαρμογές ανίχνευσης outliers περιλαμβάνουν την ανίχνευση απάτης με πιστωτικές κάρτες και την παρακολούθηση εγκληματικών δραστηριοτήτων στο ηλεκτρονικό εμπόριο κτλ.

Η ομαδοποίηση δεδομένων βρίσκεται υπό έντονη ανάπτυξη. Ως κλάδος επεξεργασίας στατιστικών στοιχείων, η ανάλυση συμπλέγματος έχει μελετηθεί εκτενώς, με κύρια εστίαση στην ανάλυση συμπλεγμάτων με βάση την έννοια της απόστασης. Εργαλεία ανάλυσης συμπλέγματος που βασίζονται στη μέθοδο «k-means» και πολλές άλλες μεθόδους έχουν επίσης ενσωματωθεί σε πολλές στατιστικές αναλύσεις ή συστήματα λογισμικού, όπως το S-Plus, το SPSS και το SAS. Στη μηχανική μάθηση, αυτή η ταξινόμηση είναι γνωστή ως εποπτευόμενη μάθηση (supervised learning) επειδή ο αλγόριθμος μάθησης εποπτεύεται (του παρέχονται κατά τη φάση εκπαίδευσης παραδείγματα σημασμένα με τις ορθές απαντήσεις). Η ομαδοποίηση είναι γνωστή ως μη εποπτευόμενη μάθηση (unsupervised learning) επειδή δεν υπάρχουν πληροφορίες ετικέτας της ορθής κλάσης (ομάδας ή συμπλέγματος) για κάθε παράδειγμα ή δεδομένο εισόδου. Για το λόγο αυτό, η ομαδοποίηση είναι μια μορφή μάθησης δια της παρατήρησης, παρά μέσω παραδειγμάτων. Στο Data Mining οι προσπάθειες έχουν επικεντρωθεί για την εύρεση μεθόδων για αποτελεσματική ανάλυση συμπλέγματος σε μεγάλες βάσεις δεδομένων **[2], [7], [10], [14], [16]**.

4.5.1. Ανάλυση Συμπλέγματος σε δεδομένα μεγάλων διαστάσεων

Οι μέθοδοι ομαδοποίησης που παρουσιάστηκαν στα προηγούμενα λειτουργούν καλά όταν οι διαστάσεις δεν είναι μεγάλες, δηλαδή το κάθε αντικείμενο αντιπροσωπεύεται από ένα διάνυσμα με από το πολύ 10 χαρακτηριστικά (σημείο στο 10-διάστατο χώρο). Υπάρχουν, ωστόσο, σημαντικές εφαρμογές όπου τα δεδομένα είναι πολύ μεγάλων διαστάσεων. Πώς μπορεί να γίνει ανάλυση συμπλέγματος για δεδομένα μεγάλων διαστάσεων; Σε αυτήν την ενότητα παρουσιάζονται τις οι διαδικασίες που συγκεντρώνουν τα δεδομένα μεγάλων διαστάσεων. Γίνεται επισκόπηση των κυριότερων προκλήσεων και των προσεγγίσεων που χρησιμοποιούνται. Οι μέθοδοι για την ενσωμάτωση δεδομένων μεγάλων διαστάσεων μπορούν να χωριστούν σε δύο κατηγορίες: ομαδοποίηση υποσυνόλων και μέθοδοι μείωσης των διαστάσεων.

Πριν παρουσιαστούν συγκεκριμένες μέθοδοι για την ομαδοποίηση δεδομένων μεγάλων διαστάσεων, εξετάζονται οι ανάγκες της ανάλυσης συμπλέγματος για τα δεδομένα μεγάλων διαστάσεων χρησιμοποιώντας παραδείγματα. Στη συνέχεια κατηγοριοποιούνται οι κύριες μέθοδοι ανάλογα με το αν αναζητούν ομοιότητες σε υποπεριοχές του αρχικού χώρου ή δημιουργούν ένα νέο χώρο με μικρότερες διαστάσεις και αναζητούν ομοιότητες εκεί.

Σε ορισμένες εφαρμογές, ένα αντικείμενο (δεδομένο) μπορεί να περιγραφεί με περισσότερα από δέκα (10) χαρακτηριστικά. Τέτοια αντικείμενα αναφέρονται ως «μεγάλος χώρος δεδομένων». Τα δεδομένα αγοράς των πελατών μίας εταιρίας, π.χ., είναι πολύ μεγάλων διαστάσεων. Επομένως, το προφίλ αγορών ενός πελάτη, που μπορεί να περιέχει τις τιμές όλων των προϊόντων που εμπορεύεται η εταιρεία, μπορεί να έχει διάσταση δεκάδων χιλιάδων. Έστω ότι εξετάζονται οι αγορές από τους πελάτες για 10 προϊόντα, έστω {P1, , P10} της εταιρίας. Εάν ένας πελάτης αγοράζει ένα προϊόν, το αντίστοιχο bit γίνεται «1», διαφορετικά γίνεται «0».

Αν υπολογιστούν οι ευκλείδειες αποστάσεις μεταξύ των πελατών, είναι εύκολο να διαπιστωθεί ότι σύμφωνα με αυτές, ενδέχεται π.χ. τρεις πελάτες να είναι ισοδύναμα όμοιοι (ή ανόμοιοι) ο ένας προς τον άλλο. Ωστόσο, μια πιο προσεκτική ματιά δείχνει ότι οι πελάτες μοιράζονται ίσως ένα κοινό προϊόν που αγοράζουν, π.χ. το P1. Ως εκ τούτου, οι ομάδες στον πλήρη, πολυδιάστατο χώρο μπορεί να είναι αναξιόπιστες, και η εύρεση τέτοιων ομάδων μπορεί να μην έχει νόημα. Τότε ποια είδη συμπλεγμάτων (ομάδων) έχουν νόημα στα δεδομένα μεγάλων διαστάσεων; Με τα συμπλέγματα ανάλυσης των δεδομένων μεγάλων διαστάσεων, στόχος είναι πάντα να ομαδοποιηθούν παρόμοια αντικείμενα (δεδομένα). Ωστόσο, ο χώρος δεδομένων είναι συχνά πολύ μεγάλος και πολύ ακατάστατος. Μια πρόσθετη πρόκληση είναι ότι πρέπει να βρεθούν όχι μόνο ομάδες, αλλά, για κάθε σύμπλεγμα, ένα σύνολο χαρακτηριστικών που αντιπροσωπεύουν το σύμπλεγμα. Με άλλα λόγια, ένα σύμπλεγμα δεδομένων υψηλών διαστάσεων συχνά ορίζεται χρησιμοποιώντας ένα μικρότερο σύνολο χαρακτηριστικών αντί για το πλήρες διάστημα δεδομένων.

Πώς μπορούν να βρεθούν συμπλέγματα από δεδομένα μεγάλων διαστάσεων; Υπάρχουν πολλές μέθοδοι. Μπορούν γενικά να κατηγοριοποιηθούν σε τρεις κύριες ομάδες:

1. οι μέθοδοι αναζήτησης υποχώρου,
2. οι μέθοδοι ενσωμάτωσης με βάση τη συσχέτιση και
3. οι μέθοδοι διύλισης.

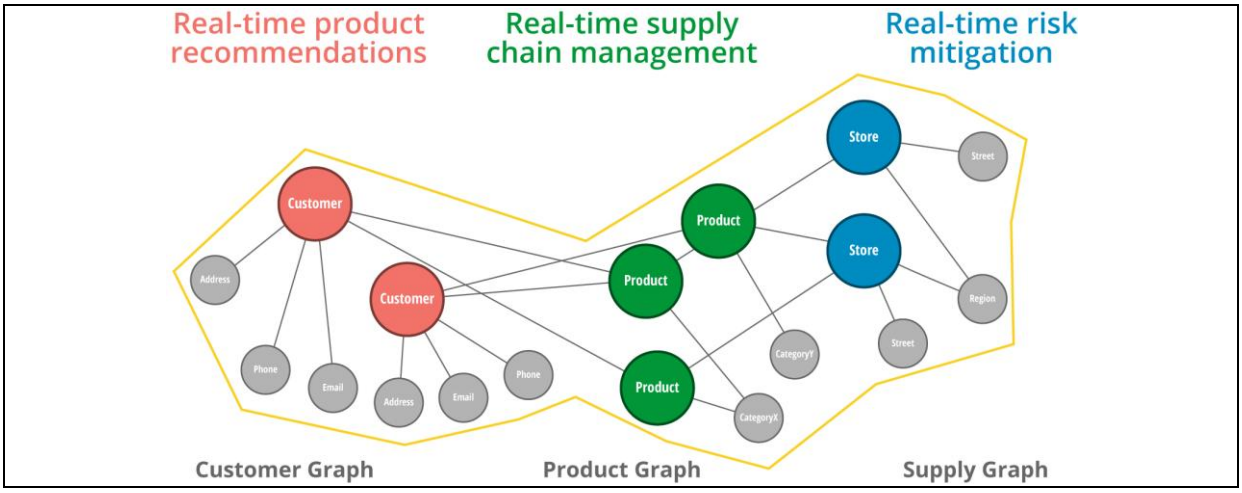
Μια μέθοδος αναζήτησης υποχώρου αναζητά συσχετίσεις μέσα σε διάφορους υποχώρους (χώρους μικρότερης διάστασης) υποσύνολα του αρχικού συνόλου δεδομένων. Εδώ, ένα σύμπλεγμα είναι ένα υποσύνολο αντικειμένων που είναι παρόμοια μεταξύ τους σε ένα υποσύστημα. Η ομοιότητα συχνά δεν μπορεί να συλληφθεί με τα συμβατικά κριτήρια (μέτρα) όπως η απόσταση ή η πυκνότητα. Για παράδειγμα ο αλγόριθμος CLIQUE είναι μια μέθοδος ομαδοποίησης υποχώρου. Απαριθμεί τις συστάδες σε μια σειρά αυξανόμενη ως προς τις διαστάσεις, και εφαρμόζει αντι-μονοτονικότητα για να απορρίπτει τους υποχώρους που δεν αντιστοιχούν σε ομάδα. Γενικά, υπάρχουν δύο είδη στρατηγικών. Οι προσεγγίσεις από κάτω προς τα πάνω που ξεκινούν από υποχώρους χαμηλής διάστασης και αναζητούν υψηλότερες διαστάσεις σε υποχώρους μόνο όταν μπορεί να υπάρχουν ομάδες σε αυτούς τους υποχώρους υψηλότερων διαστάσεων. Διάφορες τεχνικές «κλαδέματος» (pruning) διερευνώνται για τη μείωση του αριθμού των υψηλότερων διαστάσεων που πρέπει να αναζητηθούν. Το CLIQUE είναι ένα παράδειγμα προσέγγισης από κάτω προς τα πάνω. Οι προσεγγίσεις από πάνω προς τα κάτω αρχίζουν από τον πλήρη χώρο και αναζητούν αναδρομικά υποχώρους προοδευτικά όλο και χαμηλότερων διαστάσεων **[2], [7], [10], [14], [16].**

4.5.2. Ανάλυση Συμπλέγματος σε Δεδομένα Γράφου και Δικτύου

Η ανάλυση συμπλέγματος σε δεδομένα είτε οργανωμένα σε γράφο είτε σε δίκτυο, εξάγει πολύτιμες πληροφορίες. Αυτά τα δεδομένα είναι όλο και πιο δημοφιλή σε πολλές εφαρμογές όπως συγκέντρωσης γραφικών και δεδομένων δικτύου. Σημαντικό είναι να βρεθούν τα κατάλληλα μέτρα ομοιότητας για αυτή τη μορφή ομαδοποίησης.

Γενικά, οι όροι γράφος (graph) και δίκτυο (mesh ή network) μπορούν να χρησιμοποιηθούν εναλλακτικά. Γενικά, χρησιμοποιείται κυρίως ο όρος γράφος (graph). Όπως φαίνεται και στην επόμενη εικόνα αναλυτικά (Εικόνα 4.1), πολλά εταιρικά δεδομένα σχετικά με τους πελάτες και την αγοραστική συμπεριφορά τους μπορεί κατά προτίμηση να διαμορφώνονται χρησιμοποιώντας γράφους. Έτσι λοιπόν μπορεί να αναπτυχθεί ένας διμερής γράφος (bipartite graph) και η συμπεριφορά ενός πελάτη γενικότερα να απεικονιστεί σε αυτόν. Σε ένα διμερή γράφο, οι κορυφές μπορούν να χωριστούν σε δύο διαφορετικά μέρη (σύνολα) έτσι ώστε κάθε ακμή να συνδέει μια κορυφή στο ένα σύνολο με μια κορυφή στο άλλο. Τα δεδομένα αγορών πελατών, π.χ., είναι το ένα σύνολο

κορυφών που αντιπροσωπεύει τους πελάτες, με έναν πελάτη ανά κορυφή. Το άλλο σύνολο αντιπροσωπεύει τα προϊόντα, με ένα προϊόν ανά κορυφή. Μια ακμή συνδέει έναν πελάτη σε ένα προϊόν και σημαίνει την αγορά του προϊόντος από τον πελάτη **[7], [13], [16]**.



Εικόνα 4.1: Γράφημα πελάτη-προιοντος

(Πηγή. https://neo4j.com/blog/retail-neo4j-personalized-promotion-product-recommendations/)

Μηχανές αναζήτησης παγκόσμιου ιστού.

Στις μηχανές αναζήτησης παγκόσμιου ιστού, τα αρχεία καταγραφής αναζήτησης αρχειοθετούνται για την εγγραφή του χρήστη, τα ερωτήματα και τις αντίστοιχες πληροφορίες αναζήτησης ιστού. (Οι πληροφορίες αναζήτησης ιστού δείχνουν ποιες ιστοσελίδες, που δίνονται ως αποτέλεσμα μιας αναζήτησης, επισκέφθηκε ο χρήστης). Το ερώτημα και οι πληροφορίες αναζήτησης ιστού μπορούν να αντιπροσωπεύονται χρησιμοποιώντας ένα διμερή γράφο όπου τα δύο σύνολα κορυφών αντιστοιχούν σε ερωτήματα και σε ιστοσελίδες, αντίστοιχα. Μια ακμή συνδέει ένα ερώτημα σε με μία σελίδα εάν κάποιος χρήστης επισκέφθηκε την ιστοσελίδα αφού έθεσε το ερώτημα. Πολύτιμες πληροφορίες μπορούν να ληφθούν με αναλύσεις συμπλέγματος στη διμερή γραφική παράσταση ιστοσελίδων-ερωτημάτων. Για παράδειγμα, μπορεί να προσδιοριστούν ερωτήματα που τίθενται σε διαφορετικές γλώσσες, αλλά σημαίνουν ουσιαστικά το ίδιο πράγμα, εάν το οι πληροφορίες για τις επισκέψεις που ακολούθησαν μετά από το κάθε τέτοιο ερώτημα είναι παρόμοιες, **[35]**.

4.5.3. Ομαδοποίηση με περιορισμούς

Μπορεί πολλές φορές να υπάρχουν ειδικές απαιτήσεις σε μία εφαρμογή. Τέτοιες πληροφορίες μπορούν να μοντελοποιηθούν ως περιορισμοί στην ομαδοποίηση. Το θέμα της ομαδοποίησης με περιορισμούς προσεγγίζεται σε δύο βήματα. Στο πρώτο κατηγοριοποιούνται οι τύποι των περιορισμών για τη συγκέντρωση των δεδομένων σε γράφο.

Εδώ παρουσιάζονται οι τρόποι ταξινόμησης των περιορισμών που χρησιμοποιούνται στην ανάλυση συμπλέγματος. Μπορούν να κατηγοριοποιηθούν οι περιορισμοί ανάλογα με τα θέματα πάνω στα οποία έχουν τεθεί ή ανάλογα με το πόσο ισχυρά (αυστηρά) πρέπει να επιβληθούν οι περιορισμοί. Η ανάλυση συμπλέγματος περιλαμβάνει τρεις βασικές πτυχές: τα αντικείμενα ως περιπτώσεις των ομάδων, οι ομάδες ως συγκεντρώσεις των αντικειμένων και η ομοιότητα μεταξύ αντικειμένων. Επομένως, η πρώτη μέθοδος κατηγοριοποιεί τους περιορισμούς ανάλογα με τη φύση τους. Προκύπτουν έτσι τρεις τύποι περιορισμών:

- Περιορισμοί στις περιπτώσεις,
- Περιορισμοί στις ομάδες και
- Περιορισμοί στη μέτρηση ομοιότητας.

Περιορισμός στις περιπτώσεις: Ο περιορισμός στις περιπτώσεις καθορίζει τον τρόπο με τον οποίο ένα ζεύγος ή ένα σύνολο από περιπτώσεις θα πρέπει να ομαδοποιούνται στην ανάλυση συμπλέγματος.

✓ Ο **περιορισμός must-link** εάν έχει καθοριστεί σε δύο αντικείμενα x και y, τότε τα x και y πρέπει να ομαδοποιηθούν σε ένα σύμπλεγμα. Αυτοί οι περιορισμοί πρέπει να είναι μεταβατικοί. Δηλαδή, αν υπάρχουν το must-link(x, y) και το must-link(y, z) τότε πρέπει να υπάρχει το must-link(x, z).

✓ Ο **περιορισμός cannot-link**. Οι περιορισμοί που δεν μπορούν να συνδεθούν είναι το αντίθετο του must-link. Εάν ένας περιορισμός δεν μπορεί να συνδεθεί σε δύο αντικείμενα x και y, τότε στην έξοδο της ανάλυσης συμπλέγματος τα x και y θα πρέπει να ανήκουν σε διαφορετικές συστοιχίες. Αν υπάρχει ένα Cannot –link(x, y), και ένα must-link(x, x') και ένα must-link(y, y'), τότε θα υπάρχει ένα Cannot –link(x', y').

✓ **Περιορισμός στις συστάδες***:* Ο περιορισμός των συστάδων ορίζει μια απαίτηση για τις ομάδες, χρησιμοποιώντας ενδεχομένως χαρακτηριστικά των συστάδων.

Για παράδειγμα ένας περιορισμός μπορεί να καθορίσει τον ελάχιστο αριθμό αντικειμένων σε ένα σύμπλεγμα ή τη μέγιστη διάμετρο ενός συμπλέγματος ή το σχήμα ενός συμπλέγματος (π.χ. κυρτό). Ο αριθμός των ομάδων που ορίζονται για διαμέλιση ή οι μέθοδοι ομαδοποίησης μπορούν να θεωρηθούν ως ένας περιορισμός στις συστάδες.

- ✓ **Περιορισμοί στη μέτρηση ομοιότητας***:* Συχνά ένα μέτρο ομοιότητας, όπως η ευκλείδεια απόσταση χρησιμοποιείται για τη μέτρηση της ομοιότητας μεταξύ αντικειμένων στην ανάλυση συμπλέγματος. Σε ορισμένες εφαρμογές ισχύουν εξαιρέσεις. Ένας περιορισμός στη μέτρηση ομοιότητας καθορίζει την απαίτηση ότι ο υπολογισμός ομοιότητας πρέπει να τηρείται. Για παράδειγμα εάν μετρήσουμε τους ανθρώπους ως κινούμενα αντικείμενα σε μια πλατεία, η ευκλείδεια απόσταση χρησιμοποιείται για να δώσει την απόσταση περπατήματος μεταξύ δύο σημείων. Ένας περιορισμός στη μέτρηση ομοιότητας είναι ότι η τροχιά που εφαρμόζει τη μικρότερη απόσταση δεν μπορεί να διασχίσει έναν τοίχο. Μπορούν να υπάρχουν περισσότεροι από ένας τρόποι για να εκφραστεί ένας περιορισμός ανάλογα με την κατηγορία. Για παράδειγμα, μπορούμε να καθορίσουμε έναν περιορισμό στις ομάδες, π.χ. περιορισμός στη διάμετρο ενός συμπλέγματος που δεν μπορεί να είναι μεγαλύτερη από μια δεδομένη απόσταση d, **[2], [7], [10], [14], [16]**.

4.6. Συμπεράσματα

Με την ταχέως αυξανόμενη ανάπτυξη των δεδομένων σε κάθε εφαρμογή, το Data Mining πληροί την επικείμενη ανάγκη για αποτελεσματικά, κλιμακούμενα και ευέλικτα δεδομένα στην ανάλυση σε μια κοινωνία όπου η ανάγκη της γρήγορης και εξ αποστάσεως μάθησης γίνεται σιγά, σιγά αναγκαιότητα. Το Data Mining ως διαδικασία ανεύρεσης γνώσης μπορούμε να πούμε ότι μπορεί να αναλυθεί περιληπτικά στις εξής διαδικασίες:

- ✓ Τον καθαρισμό δεδομένων
- ✓ Την ενσωμάτωση δεδομένων
- ✓ Την Επιλογή δεδομένων
- ✓ Μετασχηματισμός δεδομένων
- ✓ Την ανακάλυψη προτύπων
- ✓ Την αξιολόγηση προτύπων και
- ✓ Την παρουσίαση γνώσεων.

Ένα μοτίβο είναι ενδιαφέρον εάν είναι έγκυρο σε κάποια δεδομένα με κάποιο βαθμό βεβαιότητας και καινοτομίας. Θα είναι χρήσιμο εάν μπορεί εύκολα να είναι κατανοητό από τους ανθρώπους. Μια άποψη του Data Mining αναλυμένου στις κύριες διαστάσεις του είναι:

- ✓ Δεδομένα
- ✓ Γνώση
- ✓ Τεχνολογία και
- ✓ Εφαρμογές.

Το Data Mining μπορεί να διεξαχθεί σε οποιοδήποτε είδος δεδομένων, εφόσον τα δεδομένα έχουν νόημα για μια εφαρμογή. Οι προηγμένοι τύποι δεδομένων περιλαμβάνουν χρονοσειρές ή αλληλουχίες (ακολουθίες) δεδομένων, ροές δεδομένων, χωρικά και χωροχρονικά δεδομένα, δεδομένα κειμένου και πολυμέσων, γραφήματα και δικτυωμένα δεδομένα και δεδομένα παγκόσμιου ιστού.

Η εξόρυξη δεδομένων έχει πολλές επιτυχημένες εφαρμογές, όπως επιχειρηματική ευφυΐα, ιστοεξερεύνηση, βιοπληροφορική, υγειονομική πληροφορική, χρηματοδότηση, ψηφιακές βιβλιοθήκες και ψηφιακή κυβερνήσεις. Υπάρχουν πολλά ανοικτά ζητήματα και προκλήσεις στην έρευνα περί την εξόρυξη δεδομένων. Οι ανοικτές προς διερεύνηση περιοχές περιλαμβάνουν την μεθοδολογία εξόρυξης, την αλληλεπίδραση των χρηστών, την αποτελεσματικότητα και την επεκτασιμότητα, καθώς και την αντιμετώπιση διαφόρων παραγόντων και τύπων δεδομένων. Η έρευνα για την εξόρυξη δεδομένων έχει επηρεάσει σημαντικά την κοινωνία και θα συνεχίσει να το κάνει στο μέλλον.

Όπως αναλύθηκε ήδη από την αρχή του κεφαλαίου, σημαντικό στοιχείο είναι ότι υπάρχουν πολλοί διαφορετικοί τρόποι οπτικοποίησης δεδομένων ώστε να γίνεται διευκολύνεται στο πώς βλέπει και αντιλαμβάνεται τα δεδομένα ο χρήστης. Εξίσου σημαντική είναι και η προεργασία που πρέπει να γίνει στα δεδομένα, πριν την εξόρυξη, ώστε το τελικό αποτέλεσμα να είναι το καλύτερο για το δεδομένο χρήστη μέσα στο πλαίσιο κάλυψης των αναγκών του.

Εξόρυξη και Οπτικοποίηση δεδομένων στην πλατφόρμα Moodle

5

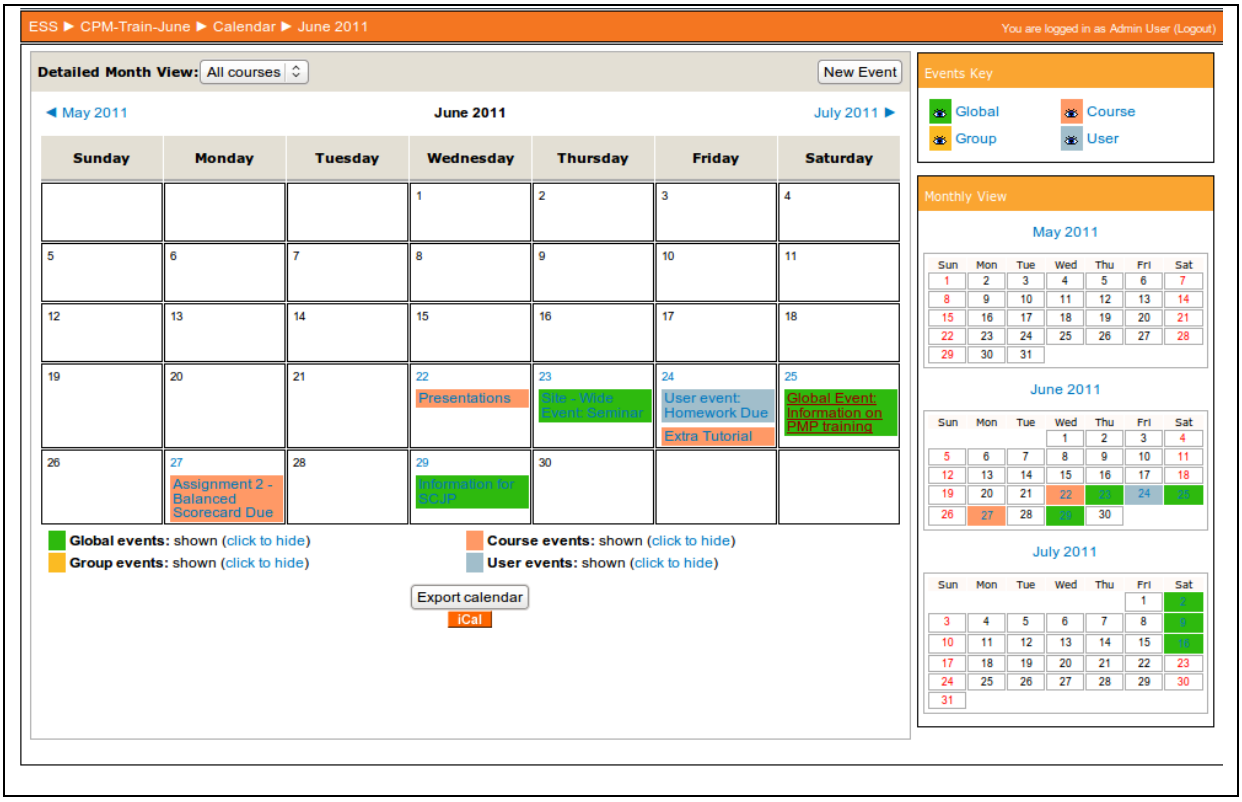
5.1. Η Πλατφόρμα Moodle

Σε αυτό το κεφάλαιο γίνεται ανάλυση των απαιτήσεων του block (τμήματος λογισμικού) που αναπτύχθηκε για τις ανάγκες της εργασίας, των εργαλείων που χρειάζονται για την ανάπτυξή του και των βημάτων για την εισαγωγή του στην πλατφόρμα moodle ώστε να εφαρμοστεί και να λειτουργήσει πάνω στη βάση δεδομένων της. Αρχικά πρέπει να διευκρινιστεί όμως τι είναι ένα block για τη moodle και που χρησιμεύει.

Γενικότερα τα blocks παρέχουν πληροφορίες και χρήσιμα εργαλεία που εμφανίζονται συνήθως στις άκρες της σελίδας στον φυλλομετρητή (web browser) του χρήστη, ή εμφανίζονται σε αναδυόμενα παράθυρα, ενώ ο χρήστης έχει ήδη συνδεθεί στην πλατφόρμα της moodle μέσω του φυλλομετρητή του.

Τα plug-ins της Moodle, που εμφανίζουν στατικές ιστοσελίδες και που υπάρχουν εκτός οποιουδήποτε «θέματος» (theme), όπως είναι για παράδειγμα οι καινούργιες ιστοσελίδες, συμπληρώνουν την πλοήγηση με το θέμα της Moodle. Ωστόσο, μπορεί επίσης να είναι ένας χρήσιμος τρόπος για να αναπτυχθεί κώδικας back-end γενικού σκοπού που δεν ταιριάζει σε κανέναν από τους άλλους τύπους plug-in. Όπως λοιπόν αναφέρθηκε, ένα plug-in είναι ουσιαστικά ένα καινούργιο περιβάλλον το όποιο έχει όλα τα χαρακτηριστικά της πλατφόρμας Moodle ενώ τα blocks είναι επεκτάσεις ή επιπλέον δυνατότητες που μπορεί να προσθέσει ο χρήστης στη Moodle. Στις επόμενες εικόνες δίνονται η μορφή ενός plug-in (Εικόνα 5.1) και ενός block (Εικόνα 5.2) για να γίνει σαφής η διαφορά, **[20], [22]**.

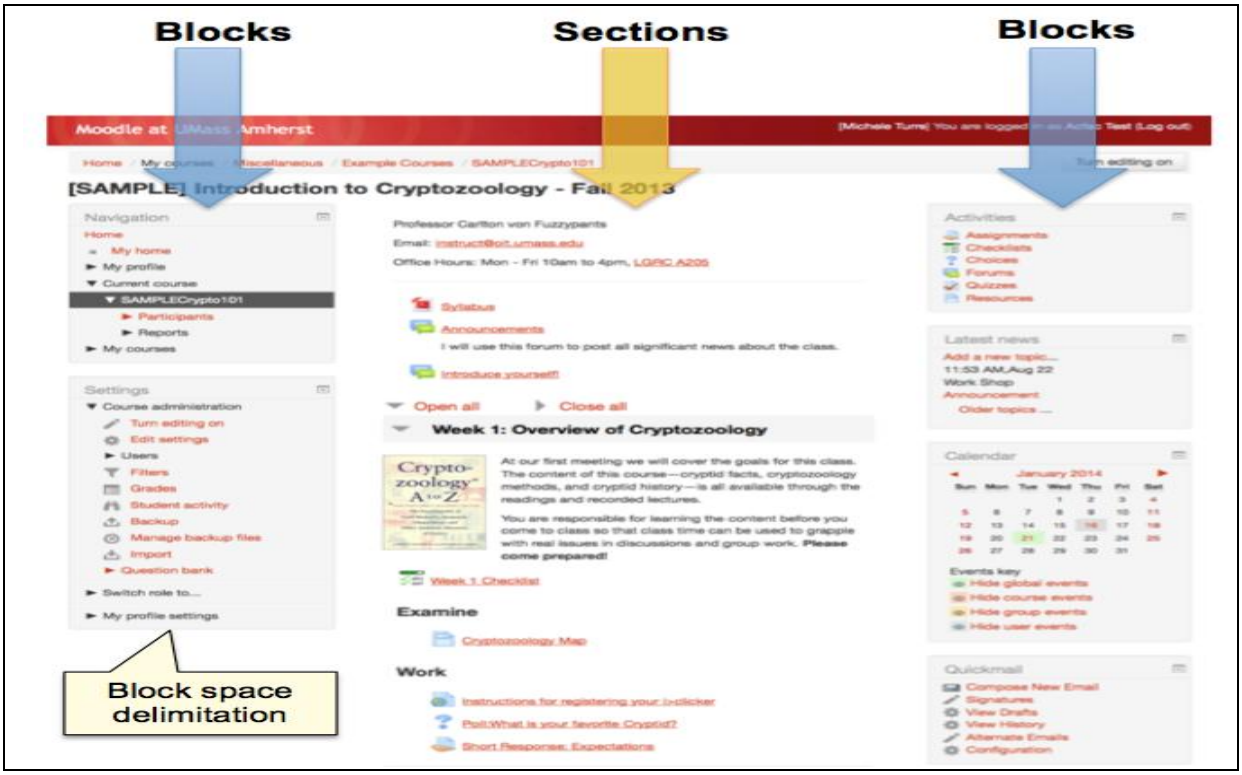
✓ Plug-in



Εικόνα 5.1: Το Plug-in «Calendar» της Moodle
(Πηγή: https://enovation.ie/one-way-syncronisation-of-your-moodle-calendar-with-outlook-or-google-calendar/)

Το αρχικό γράμμα «M» στο όνομα «Moodle» σημαίνει αρθρωτό (modular). Ο ευκολότερος και πιο συντηρητικός τρόπος προσθήκης νέων λειτουργιών στο Moodle είναι η συγγραφή ενός από αυτούς τους τύπους plug-in. Όπως είναι φανερό από την παραπάνω εικόνα, το plug-in “Calendar” έχει την δική του μορφοποίηση ανάλογα με την λειτουργία που θέλει να προσφέρει για τους χρήστες. Για παράδειγμα, το plug-in “Calendar” θα πρέπει να προσφέρει την δυνατότητα υπενθύμισης, οργάνωσης προγράμματος κτλ. ενώ αντίστοιχα θα πρέπει να προσφέρει blocks ώστε να μπορούν να εισαχθούν νέες δυνατότητες σε κάθε άλλο plug-in που υπάρχει. Ουσιαστικά αυτή είναι και η κύρια διαφορά μεταξύ plug-in και block.

✓ **Block**



Εικόνα 5.2: Blocks στη Moodle
(Πηγή: https://www.umass.edu/it/support/moodle/add-remove-blocks-moodle)

Όπως φαίνεται στην Εικόνα 5.2, τα blocks είναι στοιχεία που μπορούν να προστεθούν στην αριστερή ή τη δεξιά ή τη κεντρική στήλη (frame) οποιασδήποτε σελίδας της Moodle. Μπορούν επίσης να προστεθούν στο κεντρικό πλαίσιο του πίνακα. Εμφανίζονται στην άκρη δεξιά και είναι επιπλέον εργαλεία που μπορεί είτε ένας απλός χρήστης (καθηγητής ή φοιτητής) είτε ο διαχειριστής να χρησιμοποιήσει για να έχει επιπλέον οπτικοποίηση δυναμικών ή στατικών χαρακτηριστικών των αποτελεσμάτων του, στα πλαίσια των μαθημάτων ή των κουΐζ (π.χ. να ορίσει πώς θα βλέπει το πρόγραμμα των μαθημάτων του, να βλέπει με διαφορετική μορφή τις επιδόσεις του (ο φοιτητής) σε ένα μάθημα, ή την απόδοση των φοιτητών του (ο καθηγητής) σε ένα μάθημα, κτλ. Τα blocks εισάγονται από τον διαχειριστή της Moodle και γίνονται διαθέσιμα τόσο για τον καθηγητή όσο και για τον φοιτητή, όπως και όλες οι δυνατότητες και λειτουργίες που μπορεί να έχει αντίστοιχα ο καθένας.

5.2. PHP και Moodle

Τα προγράμματα (programs) της Moodle είναι το τμήμα εφαρμογής μιας Web εφαρμογής βάσης δεδομένων. Τα προγράμματα εκτελούν τις εργασίες που ορίζουν ο διαχειριστής ή οι χρήστες με τα κατάλληλα δικαιώματα. Κυρίως δημιουργούν και εμφανίζουν ιστοσελίδες, ώστε να δέχονται και να επεξεργάζονται πληροφορίες από τους χρήστες, να αποθηκεύουν τις πληροφορίες στη βάση δεδομένων, και να ανακτούν πληροφορίες από τη βάση δεδομένων, προκειμένου να υλοποιηθούν οι λειτουργίες ή υπηρεσίες που παρέχει η Moodle στους χρήστες της.

Η PHP είναι μια από τις γλώσσες που χρησιμοποιούνται για να συγγραφή και εισαγωγή εκτελέσιμου κώδικα στα προγράμματα της Moodle. Γενικά, είναι μια γλώσσα scripting, σχεδιασμένη για χρήση στο διαδίκτυο. Έχει χαρακτηριστικά που διευκολύνουν τον προγραμματισμό για δυναμικές εφαρμογές Web. Σε αυτήν την υποενότητα θα εξεταστούν κάποιοι γενικοί κανόνες για τη συγγραφή προγραμμάτων σε γλώσσα PHP. Οι κανόνες αυτοί θεωρούνται εξίσου βασικοί και απαραίτητοι, όπως οι κανόνες γραμματικής για τις φυσικές γλώσσες.

5.2.1. Προσθήκη κώδικα PHP σε μια σελίδα HTML

Η γλώσσα προγραμματισμού PHP είναι συνεργάτης της γλώσσας σήμανσης υπερκειμένου HTML (HyperText Markup Language). Επιτρέπει στην HTML να επιτύχει αποτελέσματα που δεν θα μπορούσε μόνη της. Για παράδειγμα, η HTML μπορεί να εμφανίσει ιστοσελίδες και η HTML διαθέτει λειτουργίες που επιτρέπουν να μορφοποιηθούν αυτές οι ιστοσελίδες. Η HTML επιτρέπει επίσης να εμφανιστούν γραφικά στις ιστοσελίδες και να αναπαραχθούν αρχεία μουσικής. Αλλά η HTML από μόνη της δεν επιτρέπει να αλληλοεπιδράσει η εφαρμογή με το χρήστη που επισκέπτεται την ιστοσελίδα. Αυτός δεν μπορεί να αποκτήσει πρόσβαση σε αυτές τις πληροφορίες χωρίς τη χρήση γλώσσας διαφορετικής από HTML. Η PHP είναι μία τέτοια γλώσσα. Επεξεργάζεται τη μορφή της πληροφορίας και επιτρέπει και άλλες παρόμοιες αλληλεπιδράσεις του χρήστη και των ιστοσελίδων που αυτός επισκέπτεται (π.χ. συμπλήρωση φόρμας ηλεκτρονικής αίτησης, κλπ.).

Οι ετικέτες (labels) HTML χρησιμοποιούνται για να κάνουν τις δηλώσεις γλώσσας PHP μέρος των HTML scripts. Το αντίστοιχο αρχείο που περιέχει τον εκτελέσιμο κώδικα μετονομάζεται με επέκταση *.php*. Στη γλώσσα PHP, οι δηλώσεις (statements) περιλαμβάνονται σε ετικέτες PHP με την ακόλουθη φόρμα:

<? php?>

Η PHP επεξεργάζεται όλες τις δηλώσεις (statements) μεταξύ δύο ετικετών PHP. Αν τα statements της PHP παράγουν αποτελέσματα, τότε το τμήμα PHP αντικαθίσταται από την έξοδο (τα αποτελέσματα). Το πρόγραμμα περιήγησης (browser) του επισκέπτη των ιστοσελίδων δεν βλέπει τον PHP κώδικα, βλέπει μόνο τα αποτελέσματά του, εάν υπάρχουν, **[4], [8]**.

5.2.2. Δημιουργία επεκτάσεων στη Moodle (Plug-ins, Blocks) με PHP

Για να δημιουργηθεί μια επέκταση (Plug-in ή Block) στη Moodle, στην πιο βασική της έκδοση, θα πρέπει να εισαχθούν μόνο τέσσερα αρχεία PHP. Ας υποτεθεί ότι θα δημιουργηθεί σε αυτό το παράδειγμα ένα *Block* που ονομάζεται 'Thesis'. Τα τέσσερα (4) αρχεία θα πρέπει να βρίσκονται στο folder Block/Thesis και είναι τα εξής:

- ✓ **1^{ο} Αρχείο: 'block_Thesis.php'**

Αυτό το αρχείο θα κρατήσει τον ορισμό της κλάσης για το *Block* '*Thesis*' (Code Snippet 5.1) και χρησιμοποιείται τόσο για να το διαχειριστεί ως πρόσθετο όσο και για να το εμφανίσει στην οθόνη. Ο κώδικας ξεκινά δημιουργώντας το κύριο αρχείο αντικειμένου, **[22]**:

```
<?php
class block_Thesis extends block_base {
    public function init() {
        $this->title = get_string('Thesis', 'block_Thesis');
    }
```

Code Snippet 5.1: Ορισμός της κλάσης
(Πηγή. https://docs.moodle.org/dev/Blocks)

Η 1^{η} γραμμή κώδικα στην εικόνα 5.1 είναι ο ορισμός κατηγορίας του Block. Πρέπει να γίνει ακριβώς με τον τρόπο που φαίνεται στην εικόνα. Μόνο το όνομα του Block (εδώ

"Thesis") μπορεί (και μάλιστα πρέπει) να αλλάξει, καθώς ορίζεται ελεύθερα κατά την επιθυμία του προγραμματιστή. Όλα τα άλλα είναι τυποποιημένα.

Στη συνέχεια (3η γραμμή στην εικόνα 5.1) δίνεται μια μικρή μέθοδος: init (). Αυτό είναι απαραίτητο για το plug-in ή το block και ο σκοπός του είναι να δώσει τιμές σε κάθε μεταβλητή-μέλος της κλάσης που χρειάζεται να δημιουργήσει παράσταση.

Σε αυτό το πολύ βασικό παράδειγμα, σκοπός της init() είναι μόνο να οριστεί this -> title, που είναι ο τίτλος που εμφανίζεται στην κεφαλίδα του block. Φυσικά ο τίτλος μπορεί να οριστεί ελεύθερα από τον προγραμματιστή. Στην 4η γραμμή στην εικόνα 5.1, έχει οριστεί να διαβάσει τον πραγματικό τίτλο από το αρχείο γλώσσας που αναφέρεται στη συνέχεια, το οποίο διανέμεται μαζί με το plug-in ή το block. Εαν απαιτείται το block να μην εμφανίζει καθόλου τίτλο, τότε θα πρέπει αυτό να οριστεί ώστε να λαμβάνει οποιαδήποτε περιγραφική τιμή επιθυμεί ο προγραμματιστής, αλλά όχι την κενή συμβολοσειρά, **[3], [22]**.

✓ **2ο αρχείο: db / access.php**

Αυτό το αρχείο θα κρατήσει τις νέες *δυνατότητες* που δημιουργούνται από το plug-in ή το block. Από την έκδοση Moodle 2.4 και μεταγενέστερα, εισήχθησαν οι δυνατότητες *addinstance* και *myaddinstance* για τον πυρήνα του block (Code Snippet 5.2). Εισήχθησαν έτσι ώστε να είναι δυνατόν να ελέγχεται η χρήση μεμονωμένων plug-ins ή blocks. Αυτές οι δυνατότητες πρέπει επίσης να προστεθούν στο προσαρμοσμένο block, οπότε το αρχείο θα είναι σε αυτή την περίπτωση στο URL: **plug-in / Thesis / db / access.php.**

Εάν το plug-in δεν πρόκειται να χρησιμοποιηθεί στη σελίδα "My Moodle", δηλαδή εάν η λειτουργία **applicable_formats** έχει οριστεί "my" ως pseudo, τότε η λειτουργία myaddinstance δεν χρειάζεται να δοθεί σε κανέναν χρήστη, αλλά πρέπει να οριστεί στο συγκεκριμένο σημείο του κώδικα, αλλιώς θα προκύψουν σφάλματα σε ορισμένες ιστοσελίδες. Στην συνέχεια παρουσιάζεται ο πίνακας *δυνατοτήτων* και ο τρόπος με τον οποίο θα πρέπει να αναζητούνται νέα plug-ins, **[9], [22]**.

```
<?php
    $capabilities = array(

    'block/Thesis:myaddinstance' => array(
        'captype' => 'write',
        'contextlevel' => CONTEXT_SYSTEM,
        'archetypes' => array(
            'user' => CAP_ALLOW
        ),
```

```
        'clonepermissionsfrom' => 'moodle/my:manageblocks'
    ),

    'block/Thesis:addinstance' => array(
        'riskbitmask' => RISK_SPAM | RISK_XSS,

        'captype' => 'write',
        'contextlevel' => CONTEXT_BLOCK,
        'archetypes' => array(
            'editingteacher' => CAP_ALLOW,
            'manager' => CAP_ALLOW
        ),

        'clonepermissionsfrom' => 'moodle/site:manageblocks'
    ),
);
```

Code Snippet 5.2: Επέκταση δυνατοτήτων του Block
(Πηγή. https://docs.moodle.org/dev/Blocks)

✓ **3ο αρχείο: lang / el / block_simplehtml.php**

Αυτό είναι το αρχείο γλώσσας για plug-ins ή blocks. Αν ο χρήστης δεν είναι αγγλόφωνος, μπορεί να αντικατασταθεί το 'en' (English) με τον κατάλληλο κωδικό γλώσσας, από τις υποστηριζόμενες από τη Moodle (άνω των 100). Όλα τα αρχεία γλώσσας για plug-ins ή blocks τοποθετούνται κάτω από τον υποφάκελο / **lang** του φακέλου εγκατάστασης του block (Code Snippet 5.3).

Από την έκδοση Moodle 2.0 και μετέπειτα, απαιτείται ένα όνομα για το block που θα εμφανίζεται στη σελίδα αναβάθμισης. Ορίζεται η τιμή του ονόματος, μαζί με τις δυνατότητες που δημιουργούνται, καθώς και άλλες γλώσσες που επιθυμεί ο χρήστης να χρησιμοποιήσει μέσα σε ένα plug-in, σε ένα πακέτο γλωσσών, όπως αναφέρθηκε προηγουμένως, **[22]**.

```
<?php
$string['pluginname'] = 'Simple HTML block';
$string['Thesis'] = 'Simple HTML';
$string['Thesis:addinstance'] = 'Add a new simple HTML block';
$string['Thesis:myaddinstance'] = 'Add a new simple HTML block to the My
Moodle page';
```

Code Snippet 5.3: Αρχεία Γλωσσών
(Πηγή. https://docs.moodle.org/dev/Blocks)

✓ **4^{ο} αρχείο: version.php**

Αυτό το αρχείο περιέχει πληροφορίες έκδοσης (version) για το plug-in ή το block μαζί με άλλες προχωρημένες παραμέτρους (Code Snippet 5.4). Πριν από την έκδοση Moodle 2.0, οι λεπτομέρειες έκδοσης για plug-ins ή τα blocks αποθηκεύονταν ως πεδία κλάσης. Από την έκδοση Moodle 2.0 και μετέπειτα, αυτά αποθηκεύονται σε ένα αρχείο που ονομάζεται **version.php**, τοποθετημένο κάτω από τον υποφάκελο **/blocks/Thesis/**. Το αρχείο έκδοσης είναι πολύ απλό: περιέχει μόνο μερικούς ορισμούς πεδίων ανάλογα με τις ανάγκες του block, **[22]**.

```
<?php
$block->component = 'block_Thesis';  $block ->version = 2011062800;  $
block ->requires = 2010112400;
```

Code Snippet 5.4: Αρχείο πληροφοριών έκδοσης
(Πηγή. https://docs.moodle.org/dev/Blocks)

Αυτό το αρχείο περιέχει ορισμούς πεδίων αντικειμένου που υποδηλώνουν το πλήρες όνομα του συστατικού frankenstyle με τη μορφή **plugintype_pluginname** και τον αριθμό έκδοσης του block μαζί με την ελάχιστη έκδοση της Moodle που πρέπει να εγκατασταθεί προκειμένου να χρησιμοποιηθεί.

Στο Plug-in που αναπτύχθηκε για τις ανάγκες της παρούσας εργασίας, τα στατιστικά στοιχεία που πρέπει να εμφανίζονται είναι τα δεδομένα από την επεξεργασία και τον υπολογισμό των βαθμών και γενικά των επιδόσεων των φοιτητών-χρηστών της πλατφόρμας. Τα δεδομένα αυτά η Moodle τα έχει τοποθετήσεις σε ξεχωριστό section, το **Moodle gradebook.** Άρα εφόσον στο πλαίσιο της εργασίας θα αναπτυχθεί ένα νέο Plug-in εντός του **Moodle gradebook,** θα πρέπει να ληφθούν υπόψη τα παρακάτω, **[37]**:

✓ Δημιουργία φακέλου **grade/report**

`grade/report/[newreport]`

✓ Δημιουργία **/db** υποφακέλου

`grade/report/[newreport]/db`

✓ Δημιουργία αρχείου **access.php** στο φάκελο db με το ακόλουθο περιεχόμενο(Code Snippet 5.5)

```
grade/report/[newreport]/db/access.php

   <?php
   $gradereport_[newreport]_capabilities = array(
       'gradereport/[newreport]:view' => array(
           'riskbitmask' => RISK_PERSONAL,
           'captype' => 'read',
           'contextlevel' => CONTEXT_COURSE,
           'legacy' => array(
               'student' => CAP_ALLOW,
               'teacher' => CAP_ALLOW,
               'editingteacher' => CAP_ALLOW,
               'admin' => CAP_ALLOW
           )
       ),
   );
   ?>
```

Code Snippet 5.5: Αρχείο access.php
(Πηγή. https://docs.moodle.org/dev/Gradebook_reports)

✓ Δημιουργία αρχείου **version.php** με τα κάτωθι data(Code Snippet 5.6):

```
grade/report/[newreport]/version.php
   <?php
   $plugin->version  = 2007081000;
   $plugin->requires = 2007081000;
   ?>
```

Code Snippet 5.6: Αρχείο version.php
(Πηγή. https://docs.moodle.org/dev/Gradebook_reports)

5.3. Data Mining στην πλατφόρμα Moodle

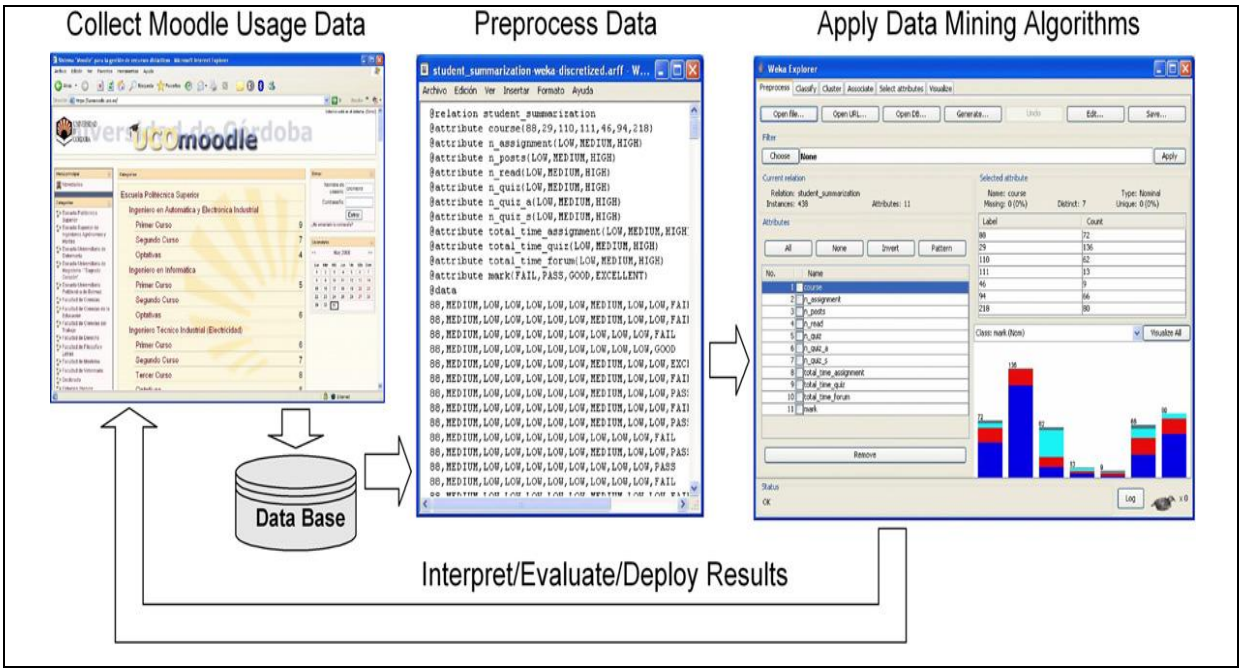
Η ανάπτυξη και εισαγωγή ψηφιακού μαθησιακού υλικού από το διδάσκοντα σε μία πλατφόρμα ηλεκτρονικής μάθησης είναι μια επίπονη και ιδιαίτερα χρονοβόρα δραστηριότητα. Ο διδάσκων ο ίδιος ή ο συνεργαζόμενος προγραμματιστής θα πρέπει να επιλέξει το περιεχόμενο που θα εμφανιστεί, να αποφασίσει για τη δομή του περιεχομένου και να καθορίσει τα καταλληλότερα πολυμεσικά στοιχεία περιεχομένου. Επιπλέον ενδεχομένως αυτά πρέπει να διαφοροποιηθούν για κάθε διαφορετικό τύπο εκπαιδευόμενου-χρήστη (μαθητή εγγεγραμμένου στο ηλεκτρονικό μάθημα), εάν υποστηρίζεται εξατομικευμένη εκπαίδευση. Λόγω της πολυπλοκότητας αυτών των αποφάσεων, είναι δύσκολο να είναι αυτές κοινές σε όλα τα ηλεκτρονικά μαθήματα που υποστηρίζει δεδομένη πλατφόρμα ηλεκτρονικής μάθησης. Η συνήθης κατάσταση είναι το κάθε ηλεκτρονικό μάθημα να έχει εσωτερικά διαφορετική δομή, οργάνωση του υλικού και

κανόνες πρόσβασης στο μαθησιακό υλικό και αξιολόγησης / βαθμολόγησης των εκπαιδευόμενων.

Από την πλευρά του προγραμματιστή, μία εφαρμογή εξόρυξης δεδομένων από ένα σύστημα ηλεκτρονικής μάθησης είναι ένας επαναληπτικός κύκλος, **[18]**. Η εξειδικευμένη γνώση πρέπει να εισέλθει στον βρόχο του συστήματος και να καθοδηγήσει, να διευκολύνει και να βελτιώσει τη μάθηση ως σύνολο, όχι μόνο μετατρέποντας τα δεδομένα σε γνώση, αλλά και φιλτράροντας τις εξειδικευμένες γνώσεις για τη λήψη αποφάσεων (Εικόνα 5.3). Η ηλεκτρονική μάθηση και η διαδικασία εξόρυξης δεδομένων από αυτήν, αποτελούνται από τα ίδια τέσσερα (4) βήματα στη γενική διεργασία εξόρυξης δεδομένων:

1. **Συλλογή δεδομένων**: Το σύστημα CMS χρησιμοποιείται από τους εκπαιδευόμενους στο πλαίσιο ενός προγράμματος σπουδών. Οι πληροφορίες πρόσβασης στο μαθησιακό υλικό, χρήσης της πλατφόρμας και αλληλεπίδρασης των χρηστών με αυτήν, αποθηκεύονται στη βάση δεδομένων της πλατφόρμας. Εδώ χρησιμοποιήθηκαν δεδομένα πρόσβασης και χρήσης της πλατφόρμας Moodle από προπτυχιακούς φοιτητές Ε εξαμήνου σπουδών, επί τρία διαδοχικά ακαδημαϊκά έτη (2015-16, 2016-17, 2017-18) του Προγράμματος Προπτυχιακών Σπουδών του Τμήματος Ηλεκτρονικών Μηχανικών του ΑΕΙ Πειραιά Τεχνολογικού Τομέα.
2. **Προεπεξεργασία των δεδομένων**: Τα δεδομένα καθαρίζονται και μετατρέπονται σε κατάλληλη μορφή για εξόρυξη. Για την προεπεξεργασία των δεδομένων της Moodle μπορεί να χρησιμοποιηθεί είτε ένα εργαλείο διαχείρισης βάσεων δεδομένων (DBMS) είτε κάποιο ειδικά σχεδιασμένο εργαλείο λογισμικού για πιο ειδική προ-επεξεργασία.
3. **Εφαρμογή εξόρυξης δεδομένων**: Οι αλγόριθμοι εξόρυξης δεδομένων εφαρμόζονται για την κατασκευή και την εκτέλεση του μοντέλου που ανακαλύπτει και συνοψίζει τις γνώσεις που ενδιαφέρουν τον χρήστη της πλατφόρμας (εκπαιδευτή/καθηγητή, εκπαιδευόμενο/φοιτητή, ή διαχειριστή). Για να το επιτευχτεί αυτό χρησιμοποιείται είτε ένα γενικό εργαλείο εξόρυξης δεδομένων είτε ένα συγκεκριμένο εργαλείο εξόρυξης δεδομένων. Μπορεί να επιλεγεί εργαλείο λογισμικού είτε εμπορικά (με χρέωση) είτε ελεύθερα διατιθέμενο.
4. **Ερμηνεία, αξιολόγηση και ανάπτυξη των αποτελεσμάτων**: Τα αποτελέσματα ή το μοντέλο που αποτελεί την έξοδο της εξόρυξης, αναλύονται, ερμηνεύονται και χρησιμοποιούνται από τον ενδιαφερόμενο χρήστη (φοιτητή / καθηγητή /

διαχειριστή) για να λάβει αποφάσεις και να σχεδιάσει και υλοποιήσει περαιτέρω ενέργειες/δράσεις βελτίωσης της εκπαίδευσης, **[15], [18], [20]**.



Εικόνα 5.3: Η Διαδικασία Data Mining στη Moodle
(Πηγή. https://www.researchgate.net/figure/Le-Cycle-dapplication-des-techniques-de-data-mining-sur-Moodle-Romero-C-Ventura_fig2_314158463)

Η εξόρυξη δεδομένων αξιοποιεί ένα πλήθος από διαφορετικές αναπαραστάσεις των αποτελεσμάτων της, όπως οι πιθανότητες, οι κανόνες συσχέτισης, τα δέντρα κλπ., καθώς και μία ποικιλία από αλγορίθμους και μεθόδους από τις περιοχές της μηχανικής μάθησης, της στατιστικής ανάλυσης και της τεχνητής νοημοσύνης.

5.4. Data Visualization στην πλατφόρμα Moodle

Η απεικόνιση πληροφοριών είναι ένας κλάδος των Γραφικών των Υπολογιστών (computer graphics) και της διεπαφής του χρήστη (user interface) που στοχεύει στη σύνθεση και προβολή κατά κανόνα 2-διάστατων ή 3-διάστατων, διαδραστικών ή κινούμενων ψηφιακών απεικονίσεων, ώστε οι χρήστες να διευκολυνθούν στην εύκολη και άμεση κατανόηση των δεδομένων. Αυτές οι τεχνικές διευκολύνουν την γρήγορη αντίληψη από το χρήστη της «ουσίας» (του μέρους της πληροφορίας που τον ενδιαφέρει) δεδομένων μεγάλης κλίμακας. Αυτό επιτυγχάνεται με την παρουσίαση των δεδομένων σε ειδική

οπτική απεικόνιση. Τυπικά, πρωτογενή δεδομένα μεγάλης κλίμακας απεικονίζονται ως γραφήματα υπολογιστικών φύλλων (spreadsheets), ως διαγράμματα ή ως 3D-απεικονίσεις.

Σε εκπαιδευτικά πλαίσια, η απεικόνιση πληροφοριών μπορεί να χρησιμοποιηθεί για τη γραφική απόδοση πολύπλοκων και πολυδιάστατων στοιχείων παρακολούθησης της προόδου φοιτητών που συλλέγονται από ηλεκτρονικά εκπαιδευτικά συστήματα, πάνω από το διαδίκτυο. Οι πληροφορίες που απεικονίζονται σε μία πλατφόρμα ηλεκτρονικής μάθησης μπορεί να αφορούν

- ✓ αναθέσεις εργασιών ή ασκήσεων στους φοιτητές,
- ✓ εισερχόμενες ερωτήσεις και απορίες προς το διδάσκοντα,
- ✓ βαθμολογίες εξετάσεων και λοιπές αξιολογήσεις,
- ✓ στοιχεία πρόσβασης και χρήσης της πλατφόρμας,
- ✓ στοιχεία προόδου στη μελέτη του ψηφιακού μαθησιακού υλικού, κλπ.

Υπάρχουν ορισμένα ειδικά εργαλεία οπτικοποίησης για τα εκπαιδευτικά δεδομένα.

- ✓ Το **CourseVis** απεικονίζει τα δεδομένα από μια διαδικτυακή πύλη java εξ αποστάσεως μέσω μιας πλατφόρμας WebCT.
- ✓ Το **GISMO** χρησιμοποιεί ως δεδομένα προέλευσης τα δεδομένα παρακολούθησης ηλεκτρονικών μαθημάτων από φοιτητές μέσω της πλατφόρμας Moodle. Δημιουργεί γραφικές παραστάσεις που μπορούν να μελετηθούν από τους εκπαιδευτές.
- ✓ Το εργαλείο **Listen** οδηγεί τους εκπαιδευτές και εκπαιδευόμενους σε αλληλεπίδραση καθώς περιηγούνται τις εκπαιδευτικές ιστοσελίδες μέσω του αυτοματοποιημένου αναγνώστη του.

Χρησιμοποιώντας αυτά τα εργαλεία, οι καθηγητές μπορούν να δουν και να μελετήσουν τις γραφικές παραστάσεις των δεδομένων που τους ενδιαφέρουν κάθε στιγμή, και που τους επιτρέπουν να κατανοήσουν την πρόοδο των φοιτητών τους και να καταλάβουν τι συμβαίνει ακόμα και σε κατηγορίες φοιτητών απομακρυσμένης πρόσβασης.

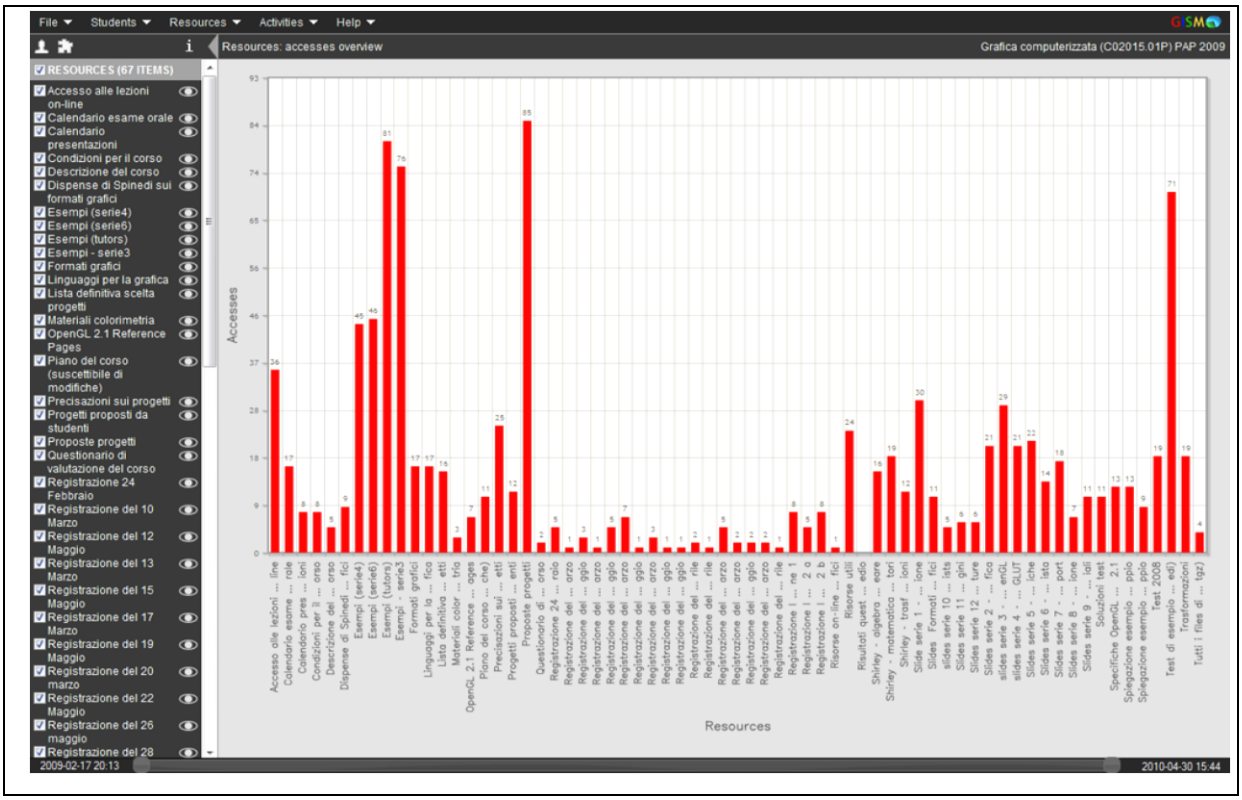
Η Moodle δεν παρέχει εργαλεία οπτικοποίησης των δεδομένων χρήσης της από τους φοιτητές, αλλά μόνο πληροφορίες κειμένου (αριθμητικές αναφορές, ανάλυση αντικειμένων, κλπ.). Μπορεί όμως να εγκατασταθεί το λογισμικό GISMO (Gismo, 2007) μέσα σε ένα σύστημα Moodle. Το GISMO είναι ένα γραφικό διαδραστικό εργαλείο παρακολούθησης φοιτητών που (α) εξάγει δεδομένα παρακολούθησης από το Moodle και (β) δημιουργεί με αυτά γραφικές παραστάσεις, που μπορούν να μελετηθούν από τους εκπαιδευτές/καθηγητές κάθε ηλεκτρονικού μαθήματος για να εξεταστούν διάφορες πτυχές

της επιτυχίας της εκπαίδευσης, σε σχέση με τη συγκεκριμένη ηλεκτρονική «τάξη» φοιτητών από απόσταση.

Το GISMO παρέχει διάφορους τύπους γραφικών παραστάσεων και αναφορών, όπως γραφήματα που αναφέρουν την πρόσβαση του μαθητή στο ηλεκτρονικό μάθημα, την πρόσβαση σε πόρους του μαθήματος, γραφικές παραστάσεις της εξέλιξης των συζητήσεων (forum) σχετικά με την αναφορά μαθημάτων και δεδομένα γραφημάτων από τα εργαλεία αξιολόγησης.

Στην επόμενη εικόνα (Εικόνα 5.4) φαίνεται ένα παράδειγμα: αναπαρίσταται ο παγκόσμιος αριθμός των προσβάσεων που πραγματοποιούνται από τους εκπαιδευόμενους (στον άξονα Χ) σε όλους τους πόρους του ηλεκτρονικού μαθήματος (άξονας Υ). Εάν ο χρήστης επιλέξει με το δεξί πλήκτρο του ποντικιού μία από τις στήλες του ιστογράμματος, με την επιλογή ''Λεπτομέρειες'' μπορεί να δει τα αναλυτικά στοιχεία πρόσβασης ενός συγκεκριμένου φοιτητή.

Χρησιμοποιώντας αυτό το γράφημα, ο εκπαιδευτής έχει μια γενική εικόνα της συνολικής πρόσβασης των φοιτητών του στο ψηφιακό μαθησιακό υλικό, με σαφή προσδιορισμό των προτύπων και τάσεων, καθώς και πληροφορίες σχετικά με τη συμμετοχή ενός συγκεκριμένου μαθητή στο μάθημα. Ξεκινώντας από αυτές τις πληροφορίες, ο εκπαιδευτής μπορεί να εντοπίσει άμεσα τους μαθητές με προβλήματα μάθησης, υστέρηση στη μελέτη της ύλης, υστέρηση στην συμμετοχή στα τεστ, μειωμένη επίδοση στις διαδικασίες αξιολόγησης, κλπ. Για παράδειγμα, φοιτητές με πολύ μικρό αριθμό προσβάσεων, πολύ μικρό αριθμό ολοκληρωμένων εργασιών/ασκήσεων ή πολύ μικρό αριθμό ολοκληρωμένων τεστ/κουίζ, μπορούν να εντοπιστεί γρήγορα και εύκολα, **[22], [35], [12]**.



Εικόνα 5.4: Απεικόνιση μέσω του εργαλείο GISMO των ενεργειών πρόσβασης (clicks) των φοιτητών στις ψηφιακές πηγές ηλεκτρονικού μαθήματος, μέσω της πλατφόρμας Moodle (Πηγή. https://docs.moodle.org/34/en/Blocks/Add-on/GISMO/overview)

5.5. Η ταξινόμηση των δεδομένων στη Moodle

Ένας ταξινομητής (classifier) είναι μια συνάρτηση απεικόνισης από ένα διακριτό ή συνεχές διάστημα χαρακτηριστικών X προς ένα διακριτό σύνολο ετικετών Y (Duda, Hart, & Stork, 2000). Η ταξινόμηση προβλέπει να έχει οριστεί συγκεκριμένος αριθμός, έστω N, διακριτών κλάσεων και να έχει αποδοθεί μία διακριτή ετικέτα (label ή tag) σε κάθε κλάση. Η εποπτευόμενη (supervised) ταξινόμηση συμβαίνει όταν

(α) σε πρώτη φάση (φάση εκπαίδευσης), ο ταξινομητής εκπαιδεύεται να ταξινομεί ορθώς ένα σύνολο γνωστών και προ-ταξινομημένων παρατηρήσεων (observations) η καθεμία από τις οποίες συνοδεύεται από την ετικέτα της κλάσης όπου πράγματι ανήκει, ενώ

(β) σε δεύτερη φάση (φάση ελέγχου), ο ταξινομητής καλείται, με βάση την εκπαίδευσή του, να ταξινομήσει ορθά μέσα στο ίδιο σύνολο κλάσεων μία νέα παρατήρηση (observation) που είναι άγνωστη (δεν έχει ετικέτα).

Στην ηλεκτρονική μάθηση, η ταξινόμηση έχει χρησιμοποιηθεί για:

- ✓ την ομαδοποίηση των φοιτητών σε ομάδες με παρόμοια χαρακτηριστικά και αντιδράσεις σε μια συγκεκριμένη παιδαγωγική στρατηγική,
- ✓ την ανίχνευση φοιτητών με συγκεκριμένο μαθησιακό πρόβλημα,
- ✓ την πρόβλεψη της απόδοσης των φοιτητών,
- ✓ την αξιολόγηση των φοιτητών,
- ✓ την ομαδοποίηση των φοιτητών σε ομάδες που μοιράζονται κοινές παρανοήσεις ή κάνουν κοινά λάθη,
- ✓ την πρόβλεψη της επίδοσης και τελικά της επιτυχίας του μαθήματος.

Το σύστημα **Keel, [36],** είναι ένα σύστημα δεδομένων που διαθέτει διάφορους αλγορίθμους ταξινόμησης. Ευρέως διαδεδομένος είναι ο αλγόριθμος **C4.5, [35]** που χρησιμοποιείται για να χαρακτηριστούν οι φοιτητές που πέρασαν ή απέτυχαν στο μάθημα. Ο C4.5 είναι ένας αλγόριθμος για τη δημιουργία δέντρων αποφάσεων (Decision Trees) και την εξαγωγή κανόνων ταξινόμησης από το δέντρο. Στην περίπτωση αυτή, ο στόχος είναι η ταξινόμηση των μαθητών σε διαφορετικές ομάδες με παρόμοια τελική βαθμολογία σε δεδομένο ηλεκτρονικό μάθημα, ανάλογα με τις δραστηριότητες που οι φοιτητές αυτοί ολοκληρώνουν μέσα στην πλατφόρμα Moodle.

Κατά την εκτέλεση του αλγορίθμου Keel, λαμβάνεται ένα δέντρο απόφασης με αριθμό κόμβων και αριθμό φύλλων στο δέντρο, αριθμό και ποσοστό σωστών και εσφαλμένα ταξινομημένων περιπτώσεων. Λαμβάνεται ένα σύνολο κανόνων του τύπου «IF-THEN-ELSE» από το δέντρο αποφάσεων που μπορεί να αναδείξει ενδιαφέρουσες πληροφορίες σχετικά με την ταξινόμηση των φοιτητών. Συγκεντρώνοντας τους κανόνες που λαμβάνονται, προκύπτουν τουλάχιστον οι εξής τρεις (3) κύριες κατηγορίες φοιτητών:

✓ Οι φοιτητές που έχουν ολοκληρώσει επιτυχώς πολύ μικρό αριθμό τεστ/κουίζ, ταξινομούνται άμεσα ως Αποτυχόντες (κλάση FAIL),

✓ Οι φοιτητές που έχουν ολοκληρώσει επιτυχώς πολύ μεγάλο αριθμό τεστ/κουίζ, ταξινομούνται άμεσα ως Εξαιρετικοί (κλάση EXCELLENT), ενώ

✓ Οι φοιτητές που έχουν ολοκληρώσει επιτυχώς έναν ενδιάμεσο αριθμό τεστ/κουίζ ταξινομούνται σε μία από τις κλάσεις {FAIL, PASS, GOOD}, ανάλογα με άλλες μετρήσεις του συστήματος (συνολικός χρόνος ενασχόλησης με τις ανατεθείσες εργασίες, αριθμός προσπαθειών και πρόσβασης στα τεστ/κουίζ, αριθμός ερωτήσεων με τις οποίες ασχολήθηκαν, αριθμός τεστ/κουίζ στα οποία προσπάθησαν αλλά απέτυχαν, κλπ.).

Ο εκπαιδευτής/καθηγητής μπορεί να χρησιμοποιήσει τις πληροφορίες και τη γνώση που του αποκαλύπτουν αυτοί οι κανόνες για τη λήψη αποφάσεων σχετικά με την αναδιαμόρφωση του ηλεκτρονικού μαθήματος για το οποίο είναι υπεύθυνος στην πλατφόρμα Moodle. Για παράδειγμα, είναι πολύ λογικό ο αριθμός των τεστ/κουίζ που ολοκλήρωσαν επιτυχώς οι περισσότεροι φοιτητές να αποτελέσει το κύριο συστατικό για την εξαγωγή της τελικής βαθμολογίας στο μάθημα. Άλλοι κανόνες μπορούν να βοηθήσουν τον εκπαιδευτή/καθηγητή να αποφασίσει την προώθηση δραστηριοτήτων ορισμένου τύπου για την επίτευξη υψηλότερων βαθμών ή αντίθετα να αποφασίσει να εξαλείψει ή να αλλάξει ριζικά εκπαιδευτικές δραστηριότητες που σχετίζονται συστηματικά με χαμηλούς βαθμούς.

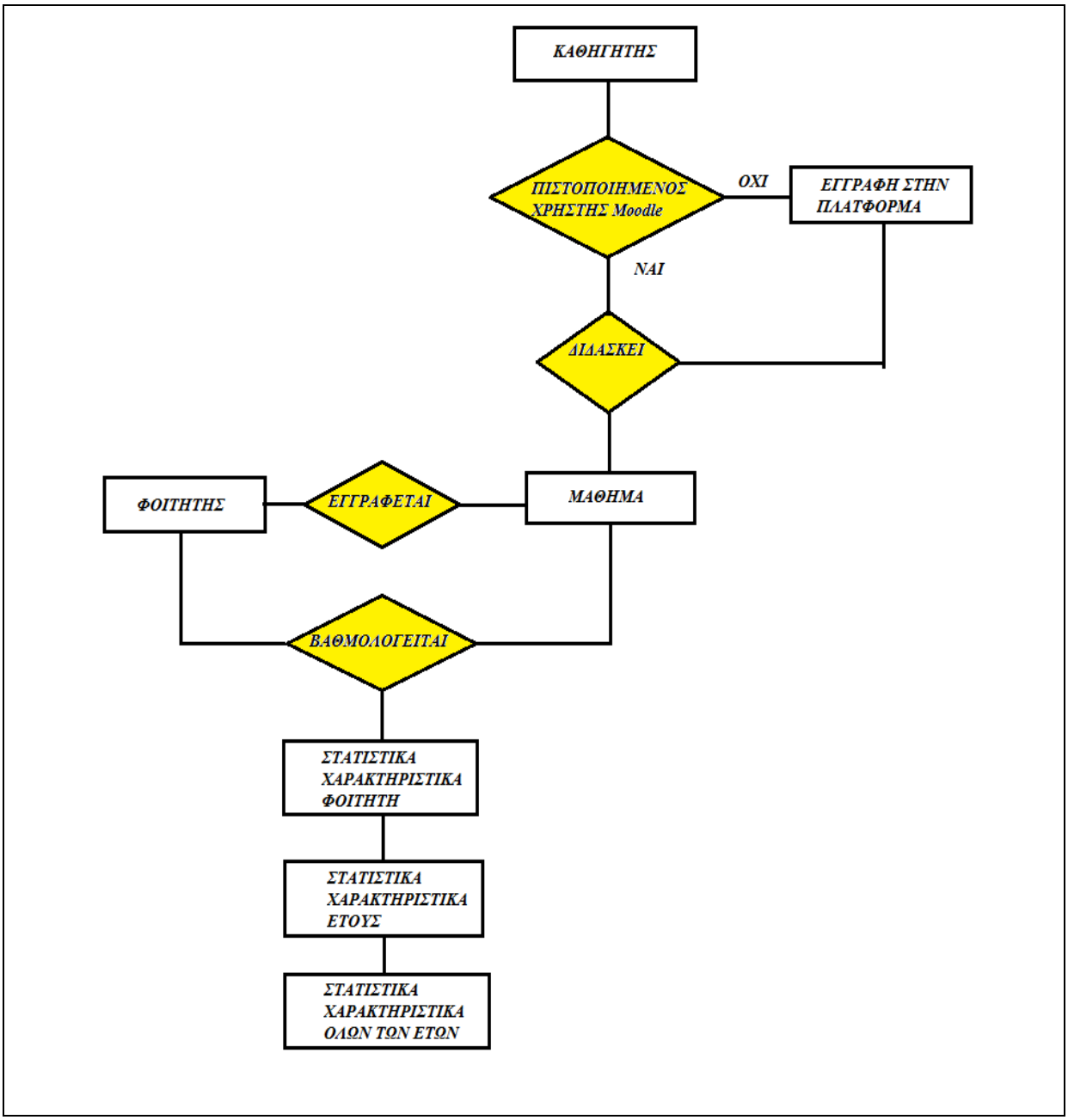
Ο εκπαιδευτής/καθηγητής μπορεί επίσης να ανιχνεύσει φοιτητές με μαθησιακά προβλήματα (φοιτητές που συστηματικά ταξινομούνται ως Αποτυχόντες (FAIL), ενώ έχουν αριθμό προσβάσεων στο σύστημα και προσπαθειών). Τέλος μπορεί να χρησιμοποιήσει το μοντέλο του δέντρου αποφάσεων για να ταξινομήσει νέο-εισερχόμενους φοιτητές και να ανιχνεύσει εγκαίρως αν θα παρουσιάσουν μαθησιακά προβλήματα κατά τη φοίτηση.

6.1. Παρουσίαση του Plug-In

Σε αυτό το κεφάλαιο παρουσιάζονται οι προδιαγραφές του Plug-In που σχεδιάστηκε και αναπτύχθηκε για τις ανάγκες της παρούσας εργασίας. Επίσης καθορίζεται η μορφή του, ώστε να γίνει κατανοητή όχι μόνο η δομή του αλλά κυρίως ο τρόπος ανάπτυξης και λειτουργίας του. Σκοπός είναι η δημιουργία μιας εφαρμογής λογισμικού που θα αποτελέσει εργαλείο για τον εκπαιδευτή/καθηγητή που χρησιμοποιεί την πλατφόρμα, όχι μόνο για τη δική του ενημέρωση και βελτίωση αλλά και για τη βελτίωση του μαθήματος και τελικά των φοιτητών του, μέσω απλών και κατανοητών στατιστικών δεδομένων για την εξαγωγή γρήγορων και κατανοητών αποτελεσμάτων.

Άλλωστε αυτή είναι και η ουσία του Data Mining (Εξόρυξης Δεδομένων) στην σημερινή πραγματικότητα, όπου η τεχνολογία πρέπει να είναι το εργαλείο διευκόλυνσης της δουλειάς των ανθρώπων. Με αυτό το σκεπτικό δημιουργήθηκε μια εφαρμογή λογισμικού όπου ο εκπαιδευτής/καθηγητής στα ηλεκτρονικά του μαθήματα θα έχει ένα πλήθος επιλογών ώστε να έχει στον υπολογιστή του γρήγορα και συγκεντρωτικά αποτελέσματα για το πλήθος των φοιτητών του αλλά και για τον καθένα μεμονωμένα. Με βάση τα εξαγόμενα στατιστικά χαρακτηριστικά θα μπορεί γρήγορα και εύκολα να βγάλει τα δικά του συμπεράσματα προς όφελος του ιδίου και της τάξης του.

Στην επόμενη εικόνα (Εικόνα 6.1) δίνεται παραστατικά η λειτουργία του Plug-In που αναπτύχθηκε, μέσω ενός Flow Diagram, ώστε να αναδειχθούν οι επιλογές που έχει ο εκπαιδευτής/καθηγητής με την εφαρμογή αυτού του Plug-In στην Moodle.



Εικόνα 6.1: Flow Diagram του Plug-in που αναπτύχθηκε.

Από το διάγραμμα ροής διαπιστώνεται ότι ο καθηγητής στο πρώτο επίπεδο επιχειρεί να εισέλθει στη Moodle. Εάν δεν είναι εγγεγραμμένος χρήστης («ΟΧΙ») θα πρέπει να εγγραφεί. Αφού εγγραφεί («ΝΑΙ») τότε θα δηλώσει το μάθημα που θα διδάξει στο οποίο μάθημα ο φοιτητής θα πρέπει να έχει εγγραφεί και για το οποίο θα βαθμολογηθεί από την καθηγητή του. Αφού βαθμολογηθεί, ο καθηγητής θα μπορεί να εξάγει συμπεράσματα μέσω των στατιστικών διαγραμμάτων της εφαρμογής με τις παρακάτω επιλογές:

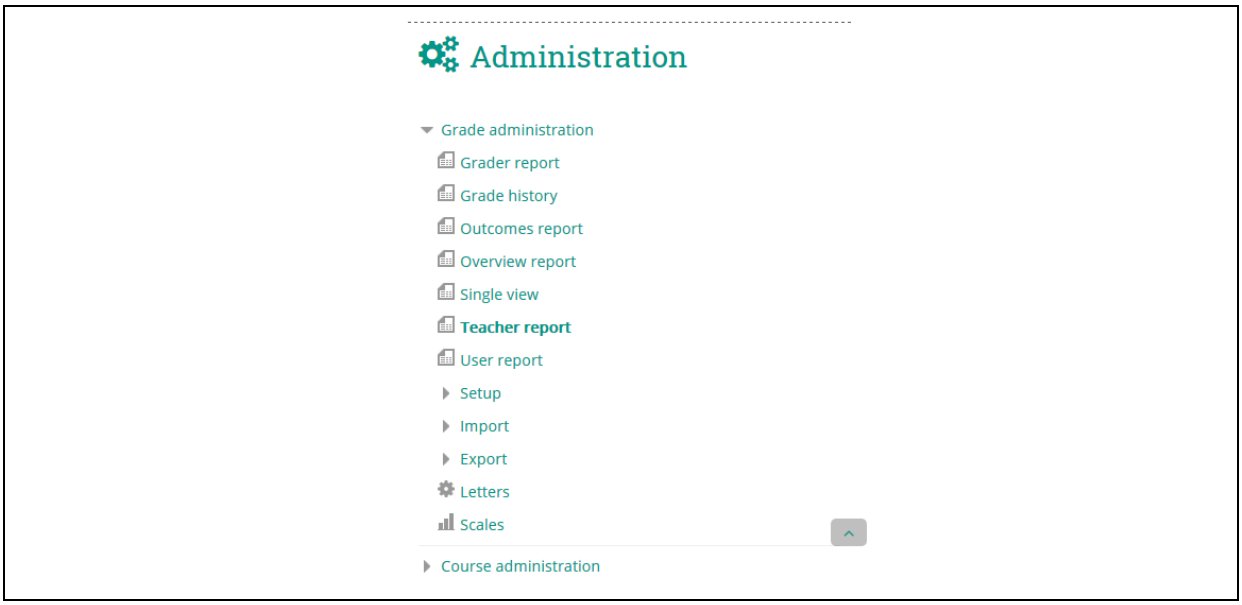
- Για την ατομική επίδοση φοιτητή,
- Για την επίδοση όλων των φοιτητών για κάθε έτος,

- Για την επίδοση όλων των φοιτητών για όλα τα έτη διδασκαλίας του μαθήματος.

Αυτές τις απεικονίσεις θα τις βλέπει όχι μόνο με γραφήματα **Bar Chart** για κάθε φοιτητή ή για όλα τα έτη, αλλά και με στατιστικά στοιχεία για τον κάθε φοιτητή χωριστά σε σύγκριση με το μέσο όρο της τάξης. Στην επόμενη ενότητα δίνεται αναλυτικά η απεικόνιση που θα βλέπει ο καθηγητής και για τις τρεις αυτές περιπτώσεις που προαναφέρθηκαν.

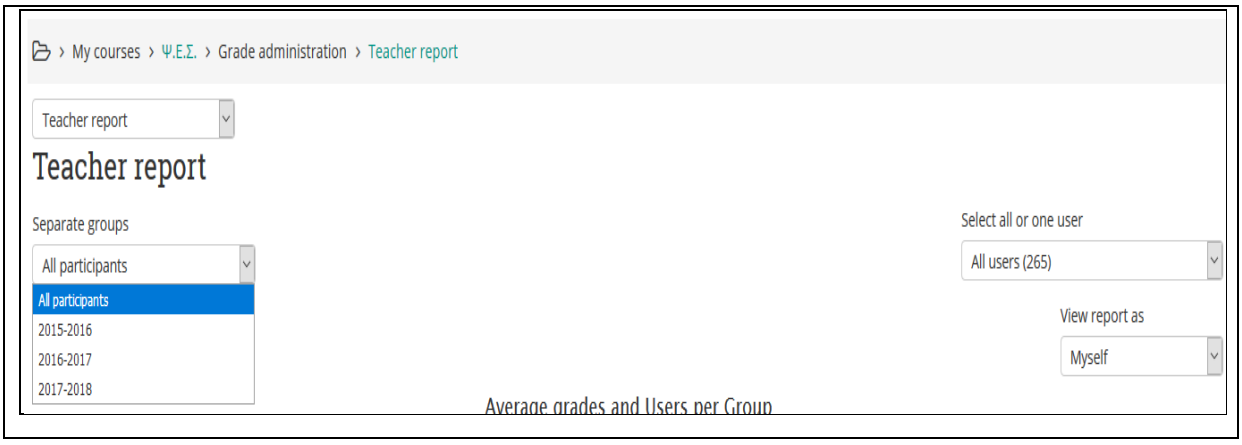
Το plug-in που σχεδιάστηκε και αναπτύχθηκε ονομάζεται **Teacher Report** και μπορεί ο διαχειριστής της Moodle να το εγκαταστήσει με τον συνήθη τρόπο εγκατάστασης των plug-ins της Moodle. Το **Teacher Report** είναι επέκταση του υπάρχοντος εντός της Moodle plug-in **Grade/Report.** Ο λόγος που επιλέχθηκε να γίνει αυτό είναι πολύ σημαντικός και ουσιώδης: έτσι ενσωματώνονται πλήρως στο νέο plug-in όλες οι δυνατότητες του **Gradebook** της Moodle, μέσα στο οποίο βρίσκονται τα plug-ins **Grade/Report**. Έτσι δεν χρειάζεται να αναπτυχθεί κώδικας από μηδενική βάση για όλα τα επιθυμητά χαρακτηριστικά του νέου plug-in, καθώς τα περισσότερα εισάγονται (κληρονομούνται) από το **Gradebook** όπου ανήκει το προϋπάρχον plug-in **Grade/Report.**

Μετά την εγκατάσταση του νέου plug-in **Teacher Report**, ο καθηγητής έχει την δυνατότητα να επιλέξει μέσω της επιλογής **Teacher Report (**που υπάρχει στο **Gradebook setup)** σε ποιό από τα μαθήματά του θέλει να εφαρμόσει τις δυνατότητες του νέου plug-in (Εικόνα 6.2).



Εικόνα 6.2: Επιλογή του Plug-in “Teacher Report” από το Gradebook Setup.

Αυτό το επιτυγχάνει επιλέγοντας κατάλληλα από το **Drop down menu** την απεικόνιση που θέλει. Όπως φαίνεται στην συνέχεια (Εικόνα 6.3), επάνω αριστερά ο καθηγητής επιλέγει το όνομα της αναφοράς που θέλει να δει (το νέο plug-in ή κάποιο άλλο από τα διαθέσιμα reports), κάτω αριστερά επιλέγει το ακαδημαϊκό έτος ή όλα τα έτη, επάνω δεξιά επιλέγει όλους τους φοιτητές ή έναν μεμονωμένο φοιτητή που θέλει να δει και κάτω δεξιά επιλέγει εάν θέλει να δει το Teacher Report ως καθηγητής ή ως διαχειριστής. Εδώ πρέπει να τονιστεί ότι η διαφορά των ρόλων καθηγητή και διαχειριστή δίνει ή αποκλείει την πρόσβαση σε ακριβώς αυτό το **Drop down menu.** Όταν ο χρήστης εισέλθει στην εφαρμογή με δικαιώματα καθηγητή, απλά δεν έχει την επιλογή αυτή. Αντίθετα, ο διαχειριστής μπορεί να επιλέξει να δει τα γραφήματα είτε ως καθηγητής και ως διαχειριστής, αφού έχει το υπερσύνολο των δικαιωμάτων πρόσβασης. Όμως αυτό δεν αλλάζει τίποτα στην απεικόνιση των δεδομένων, όπως θα αναλυθεί στη συνέχεια. Ο διαχειριστής (admin) και ο καθηγητής βλέπουν ακριβώς τα ίδια πράγματα στην απεικόνιση.



Εικόνα 6.3: Επιλογές του Plug-in Teacher Report, με τα τέσσερα (4) Drop-down menus.

Ο φάκελος του Plug-in περιέχει μέσα τα τέσσερα αρχεία όπου αναφέρθηκαν στο Κεφάλαιο 5 τα οποία είναι απαραίτητα για την λειτουργία του plug-in και την συμβατότητά του με την πλατφόρμα Moodle ενώ αναλυτικά ο κώδικας εμφανίζεται στο **Παράρτημα 1**. Επίσης περιέχει έναν υπο-φάκελο **.js** όπου έχουν εισαχθεί οι συναρτήσεις (**functions)** και ορίζεται πώς θα εκτελούνται, δηλαδή τι και πώς θα φαίνεται στην απεικόνιση των δεδομένων που καλούνται από το βασικό αρχείο **Index** του Plug-in. Επίσης εκεί βρίσκεται και το αρχείο **teacher_report.php** όπου έχουν γραφτεί οι κλάσεις των μεθόδων της js δηλαδή οι συνθήκες και ο τρόπος που θα εκτελούνται οι μέθοδοι. Όπως θα γίνει φανερό

στα επόμενα, η βιβλιοθήκη έτοιμων συναρτήσεων που καλείται για την οπτικοποίηση των δεδομένων από το νέο Plug-in είναι η **Highcharts.**

6.2. Ανάλυση λειτουργίας του κώδικα του plug-in

Όπως έχει αναφερθεί στην προηγούμενη ενότητα, εφόσον το νέο plug-in είναι μέρος του **Gradebook** της Moodle και επέκταση του **Grade/Report**, τα αρχεία του είναι ιδία με εκείνου, εκτός από τα δύο (2) αρχεία που προαναφέρθηκαν και από ορισμένες μεθόδους του αρχείου **Index,** του βασικού δηλαδή αρχείου που διαβάζει το plug-in. Το αρχείο **js** είναι το αρχείο όπου έχουν γραφτεί οι μέθοδοι και το αρχείο που χρησιμοποιεί τη βιβλιοθήκη της οπτικοποίησης για απεικόνιση. Όλες οι μέθοδοι λειτουργούν με τον ίδιο τρόπο, ως εξής:

```
function groupsAverageHChart (allGroupNamesArray, countGroupUsersArray,
countGroupGradedUsersArray, averageGroupGradeArray) {

    Highcharts.chart('allGroupsHChart', {
    chart: {
        zoomType: 'xy'
    },
    title: {
        text: 'Average grades and Users per Group'
    },
    subtitle: {
        text: ''
    },
    xAxis: [{
        categories: allGroupNamesArray,
        crosshair: true
    }],
    yAxis: [{ // Primary yAxis
        labels: {
        format: '{value}',
        style: {
            color: Highcharts.getOptions().colors[1]
        }
        },
        title: {
```

```
            text: 'Average grade',
            style: {
                color: Highcharts.getOptions().colors[1]
            }
        },
        min: 0,
        max: 10,
        allowDecimals: false,
    }, { // Secondary yAxis
        title: {
        text: 'Users',
        style: {
            color: Highcharts.getOptions().colors[0]
        }
        },
        labels: {
        format: '{value}',
        style: {
            color: Highcharts.getOptions().colors[0]
        }
        },
        opposite: true
    }],
    tooltip: {
        shared: true
    },
    series: [{
            name: 'Graded users',
            type: 'column',
            yAxis: 1,
            data: countGroupGradedUsersArray,
            tooltip: {
            valueSuffix: ''
            },
            color: Highcharts.getOptions().colors[2]
        }, {
            name: 'Users',
            type: 'column',
            yAxis: 1,
            data: countGroupUsersArray,
            tooltip: {
```

```
            valueSuffix: ''
            },
            color: Highcharts.getOptions().colors[0]
        }, {
            name: 'Average grade',
            type: 'spline',
            data: averageGroupGradeArray,
            tooltip: {
            valueSuffix: ''
            },
            color: Highcharts.getOptions().colors[1]
        }]
    });
}
```

Στην 1η γραμμή του κώδικα ορίζεται μια μέθοδος **function** με το όνομα **groupsAverageHChart** μέσα στην οποία δημιουργούνται τέσσερις μεταβλητές. Οι μεταβλητές αυτές δεν εξαρτώνται από κάποιο άλλο αρχείο και είναι μεταβλητές που στην **js** δεν απαιτούν ορισμό. Στην επόμενη γραμμή χρησιμοποιείται η γραφική **chart** από την βιβλιοθήκη **Highcharts**. Ακριβώς από κάτω ορίζονται τα χαρακτηριστικά της (title, margin, subtitle κτλ) και εάν θα είναι 3-D ή όχι, στοιχείο που ορίζεται από το true or false του enable.

Στη συνέχεια δίνεται ένας τίτλος στη γραφική παράσταση και ορίζονται οι ρυθμίσεις για τη χάραξη (plot). Έπειτα ο κώδικας προχωρά ρυθμίζοντας για τους άξονες X και Y το στιλ τους και ποιά μεταβλητή θα απεικονίζουν. Για παράδειγμα, ο άξονας Y μπορεί να επιστρέφει στην μέθοδο δύο μεταβλητές από κάθε user, και value που δηλώνεται αμέσως από κάτω για τον ίδιο άξονα. Αμέσως πιο κάτω στον κώδικα ρυθμίζεται πώς θα ομαδοποιηθούν αυτές οι μεταβλητές που επιστρέφει η μέθοδος.

Στη συνέχεια, με την εντολή **tooltip** ρυθμίζεται το σημείο όπου θα εμφανίζονται τα στοιχεία για την γραφική με όλα τα χαρακτηριστικά της (μέγεθος, χρώμα κτλ). Με ανάλογο τρόπο λειτουργούν όλες οι functions του αρχείου **js.**

Οι μεταβλητές όλων των μεθόδων του αρχείου καλούνται όπως και οι μέθοδοι από την **Index,** το βασικό αρχείο του plug-in, εκεί όπου οι μεταβλητές από array που έχουν δημιουργήσει τα αρχεία θα επιστραφούν ως string. Αυτό συμβαίνει διότι η **js** είναι μια γλώσσα που εκτελείται στον browser και δεν μπορεί να απεικονίσει arrays. Φυσικά

υπάρχει και ο εναλλακτικός τρόπος να εισαχθεί καθαρός κώδικας js μέσα στο αρχείο php, όμως έτσι θα είχε δημιουργηθεί ένα δυσνόητο και δύσκολο αρχείο.

Στη συνέχεια παρουσιάζεται μια κλάση από το αρχείο **teacher_report.php** και αναλύεται η λειτουργία της.

```
class teacher_report_group_user_data {
    public $id="";
    public $name="";
    public $grade = 0;
    public function __construct($userId, $userName, $userGrade = 0) {
        $this->id = $userId;
        $this->name = $userName;
        $this->grade = $userGrade;
    }
}
```

Σε αυτό το αρχείο γίνεται ουσιαστικά η ομαδοποίηση των δεδομένων ώστε να μπορούν να ανακτώνται με περισσότερη ευκολία αλλά και με μαζικότερο τρόπο. Γενικά αυτό το αρχείο έχει 3 κλάσεις, την **group_user_data,** την **group_data** και την **all_groups_data**.

- ✓ Η user_data είναι η μεταβλητή η οποία δημιουργεί ένα πίνακα μέσα στον οποίο καταχωρεί τα στοιχεία του κάθε user.
- ✓ Η group_data είναι η θέση ενός πίνακα που αποθηκεύονται τα στοιχεία ενός group που μπορεί να έχει πολλούς users.
- ✓ Τέλος, η all_groups_data είναι ο πίνακας που αποθηκεύει τα δεδομένα όλων των groups. Η μέθοδος που αναλύεται στη συνέχεια είναι η group_data.

Στην αρχή ορίζεται η κλάση και δηλώνεται το όνομά της. Έπειτα δίνεται εντολή στην κλάση να «τραβήξει» τα στοιχεία { **id**, **name, grade** } των φοιτητών και μαζί με τα δεδομένα της κλάσης **user_data** που έχουν δηλωθεί πιο πάνω, ξεκινώντας να δημιουργήσει έναν πίνακα που θα δώσει ένα ατομικό { **Id, Name, Grade** }. Στο αρχείο αυτό, το σύμβολο $ σημαίνει ότι η μεταβλητή που δηλώνεται είναι μόνιμη. Αυτά τα στοιχεία θα εξαχθούν αργότερα από την Index.

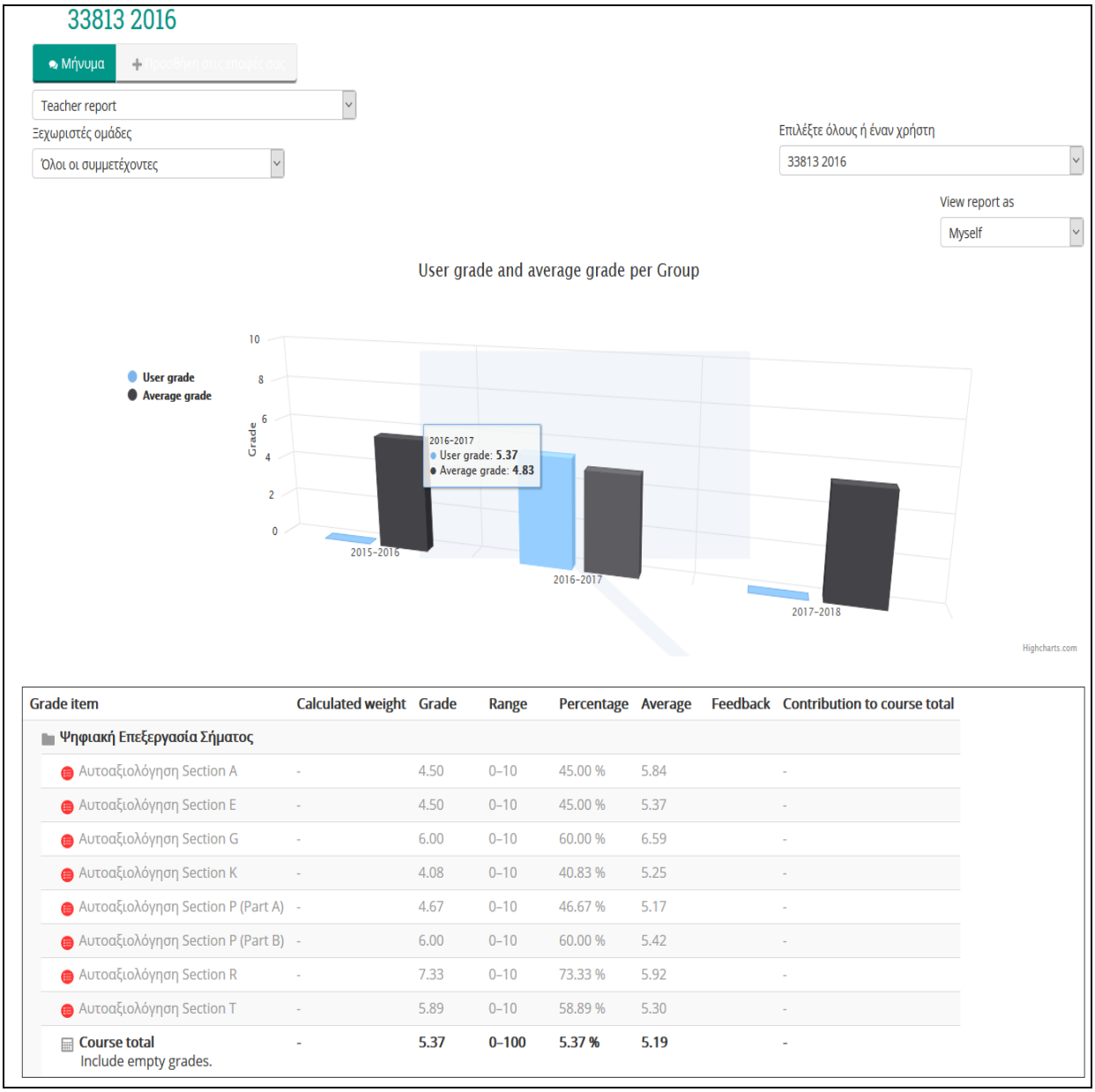
6.3. Αποτελέσματα εφαρμογής του Plug-in στη βάση δεδομένων Moodle

Για να παρουσιαστούν και να αξιολογηθούν τα αποτελέσματα της εφαρμογής του plug-in Teacher Report, χρησιμοποιήθηκαν ως βάση τα πραγματικά στοιχεία από την πραγματική βάση του εξυπηρετητή Moodle του Τμήματος Ηλεκτρονικών Μηχανικών του (πρώην) ΑΕΙ Πειραιά Τεχνολογικού Τομέα, όπως αυτά συγκεντρώθηκαν από την εκπαιδευτική διαδικασία του υποχρεωτικού προπτυχιακού μαθήματος 5^{ου} εξαμήνου «Ψηφιακή Επεξεργασία Σήματος» (Ψ.Ε.Σ.) κατά τα τρία (3) διαδοχικά ακαδημαϊκά έτη 2015-16, 2016-17 και 2017-18. Οι βαθμολογίες είναι όλες στην κλίμακα 0 – 10.

Εξετάζεται η περίπτωση χρήστη που συνδέεται στην πλατφόρμα Moodle με τα δικαιώματα του διαχειριστή (admin) – οπότε θα έχει την ίδια απεικόνιση με τον καθηγητή. Ο χρήστης-διαχειριστής μπορεί να δει τα όλα τα μαθήματα στα οποία είναι διαχειριστής, ενώ επίσης μπορεί με τα δικαιώματα που έχει στη Moodle να αναρτά μαθησιακό υλικό, τεστ, αναθέσεις εργασιών, κλπ. Η εμφάνιση των στατιστικών στοιχείων μετά την εγκατάσταση του νέου plug-in, για παράδειγμα, στα διάφορα μαθήματα που έχει ο χρήστης δικαιώματα διαχειριστή, είναι η εξής:

- **Ατομική επίδοση φοιτητή**

Χρησιμοποιείται ως παράδειγμα το μάθημα «Ψηφιακή Επεξεργασία Σήματος» (Ψ.Ε.Σ.) και ο φοιτητής με Αριθμό Μητρώου ID 33813. Το αποτέλεσμα λειτουργίας του plug-in είναι η εμφάνιση των εξής γραφημάτων:



Grade item	Calculated weight	Grade	Range	Percentage	Average	Feedback	Contribution to course total
Ψηφιακή Επεξεργασία Σήματος							
Αυτοαξιολόγηση Section A	-	4.50	0–10	45.00 %	5.84		-
Αυτοαξιολόγηση Section E	-	4.50	0–10	45.00 %	5.37		-
Αυτοαξιολόγηση Section G	-	6.00	0–10	60.00 %	6.59		-
Αυτοαξιολόγηση Section K	-	4.08	0–10	40.83 %	5.25		-
Αυτοαξιολόγηση Section P (Part A)	-	4.67	0–10	46.67 %	5.17		-
Αυτοαξιολόγηση Section P (Part B)	-	6.00	0–10	60.00 %	5.42		-
Αυτοαξιολόγηση Section R	-	7.33	0–10	73.33 %	5.92		-
Αυτοαξιολόγηση Section T	-	5.89	0–10	58.89 %	5.30		-
Course total Include empty grades.	-	**5.37**	**0–100**	**5.37 %**	**5.19**		-

Εικόνα 6.4: User ID 33813: Ατομικός βαθμός και μέσος βαθμός κατά τη διάρκεια του ακαδημαϊκού έτους, σε όλα τα ακαδημαϊκά έτη.

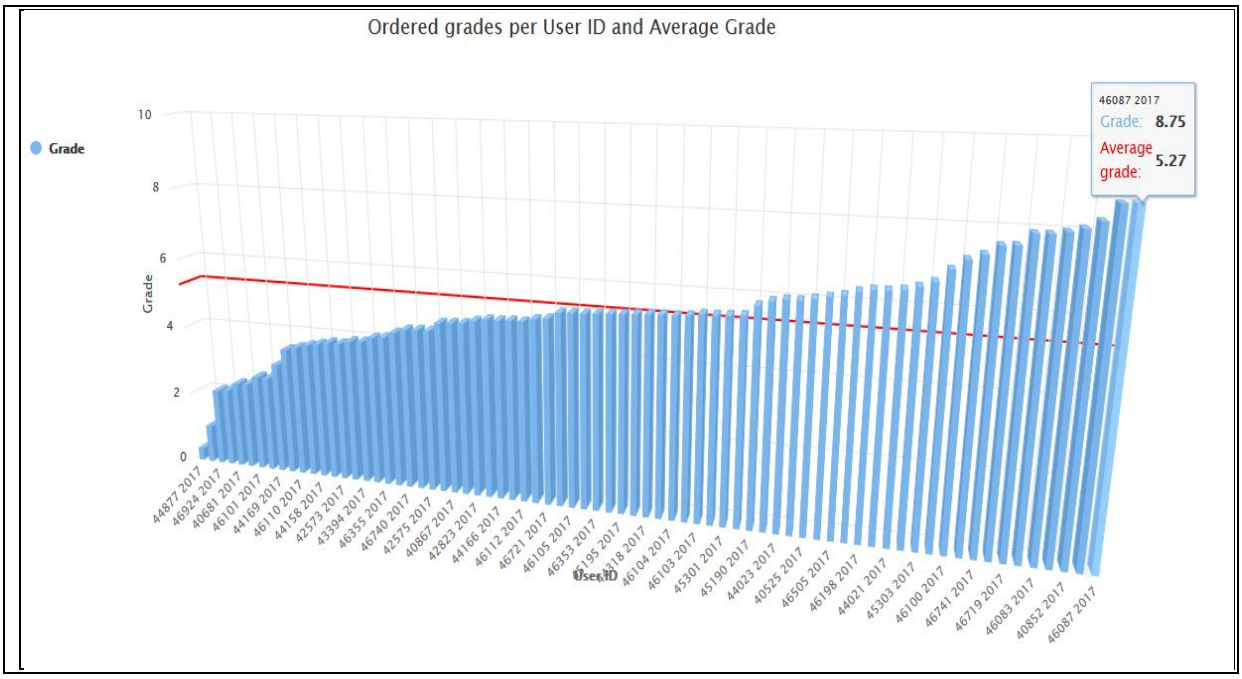
Στην Εικόνα 6.4 φαίνεται σε τρισδιάστατη απεικόνιση η ατομική απόδοση του φοιτητή σε σχέση με τον μέσο όρο της τάξης του και απεικονίζεται με bar chart. Όπως φαίνεται απεικονίζονται όλα τα έτη που ο φοιτητής είτε προσπάθησε και απέτυχε είτε ήταν εγγεγραμμένος και δεν παρακολούθησε το μάθημα. Στο κάτω μέρος της Εικόνας 6.3, δίνονται αναλυτικά οι βαθμολογίες του φοιτητή και οι μέσοι όροι της τάξης του, σε όλα τα ενδιάμεσα τεστ.

- **Επίδοση όλων των φοιτητών για κάθε έτος**

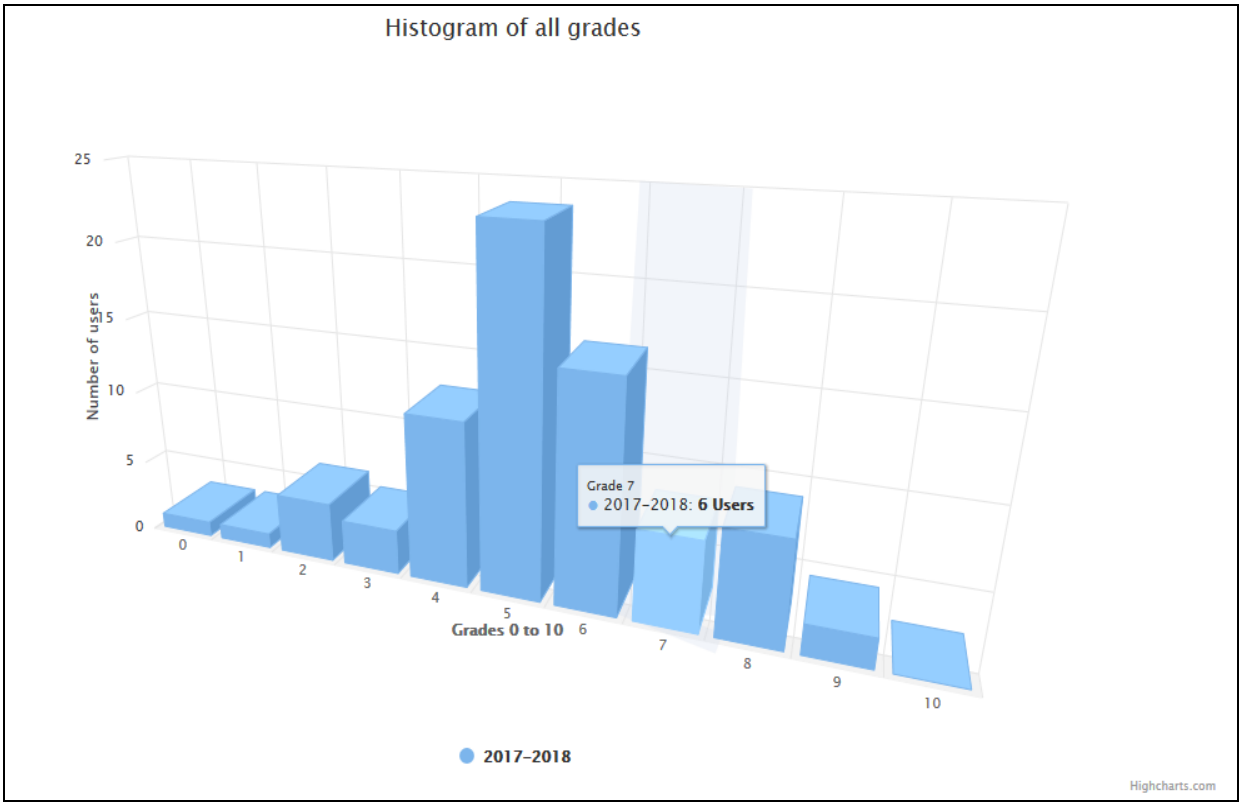
Για το ίδιο μάθημα (Ψ.Ε.Σ.) εμφανίζεται η συνολική στατιστική εικόνα της τάξης, με δυο τρισδιάστατες γραφικές.

Η πρώτη Εικόνα 6.5 απεικονίζει τους ατομικούς τελικούς βαθμούς στο συγκεκριμένο μάθημα, για όλους τους εγγεγραμμένους στο μάθημα αυτό φοιτητές. Ο κάθε φοιτητής αποδίδεται με μια κατακόρυφη στήλη. Έχουν ταξινομηθεί από τη χαμηλότερη προς την υψηλότερη βαθμολογία. Περνώντας το ποντίκι πάνω από συγκεκριμένη στήλη, εμφανίζονται σε ορθογώνιο πλαίσιο τα ακριβή στοιχεία του συγκεκριμένου φοιτητή που αντιστοιχεί σε αυτή τη στήλη, σε σχέση με την τάξη του. Για παράδειγμα, στην τελευταία δεξιά στήλη της Εικόνας 6.5, εμφανίζεται σε πλαίσιο για το φοιτητή με ID 46087-2017 ο τελικός βαθμός (8,75 με άριστα το 10,00) και ο μέσος όρος της τάξης του (5,27 με άριστα το 10,00). Ο μέσος όρος αποδίδεται οπτικά με κόκκινη οριζόντια γραμμή στο κατάλληλο ύψος.

Η επόμενη Εικόνα 6.6 απεικονίζει σε κλίμακα 0-10 τον αριθμό των φοιτητών του έτους που επέτυχε την αντίστοιχη βαθμολογία. Και εδώ περνώντας το ποντίκι πάνω από συγκεκριμένη στήλη βαθμού (βαθμός οκτώ (8) για παράδειγμα στην Εικόνα 6.4), εμφανίζονται σε ορθογώνιο πλαίσιο τα ακριβή στοιχεία για τη συγκεκριμένη βαθμολογία (επτά (7) φοιτητές έλαβαν τη συγκεκριμένη βαθμολογία ως τελικό βαθμό μαθήματος, κατά το έτος 2017-18).



Εικόνα 6.5: Επιλεγμένες βαθμολογίες ανά αναγνωριστικό χρήστη και μέση βαθμολογία κατά το ακαδημαϊκό έτος 2017-18

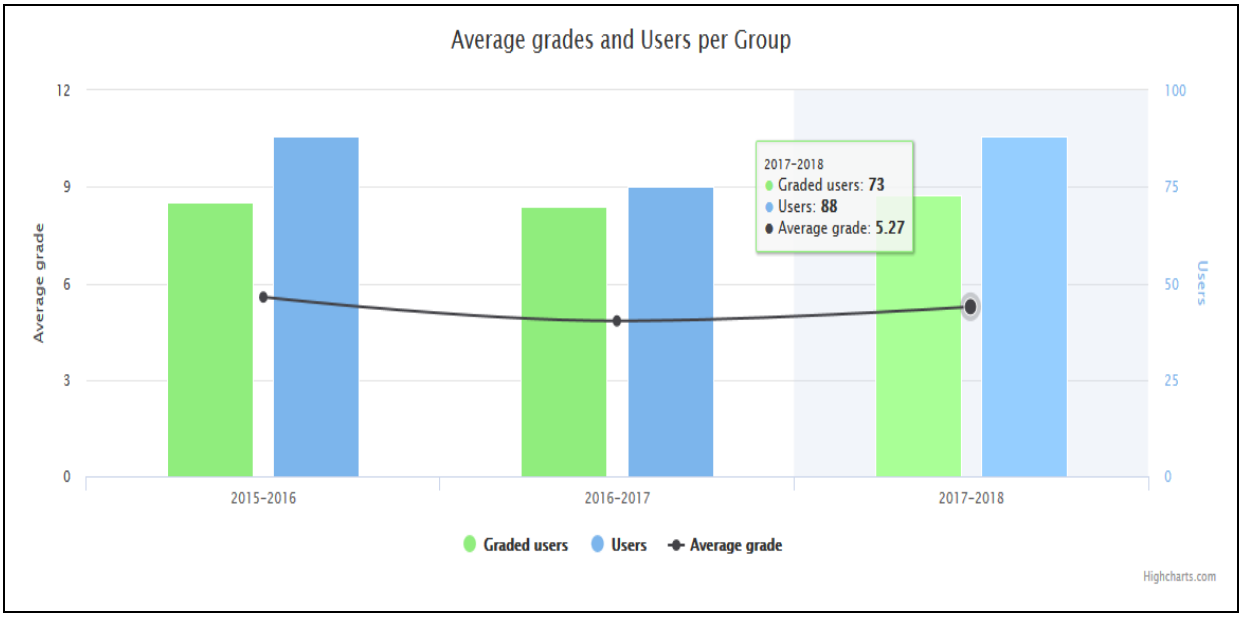


Εικόνα 6.6: Ιστόγραμμα των τελικών βαθμολογιών κατά το ακαδημαϊκό έτος 2017-18.

- **Επίδοση όλων των φοιτητών για όλα τα έτη διδασκαλίας του μαθήματος**

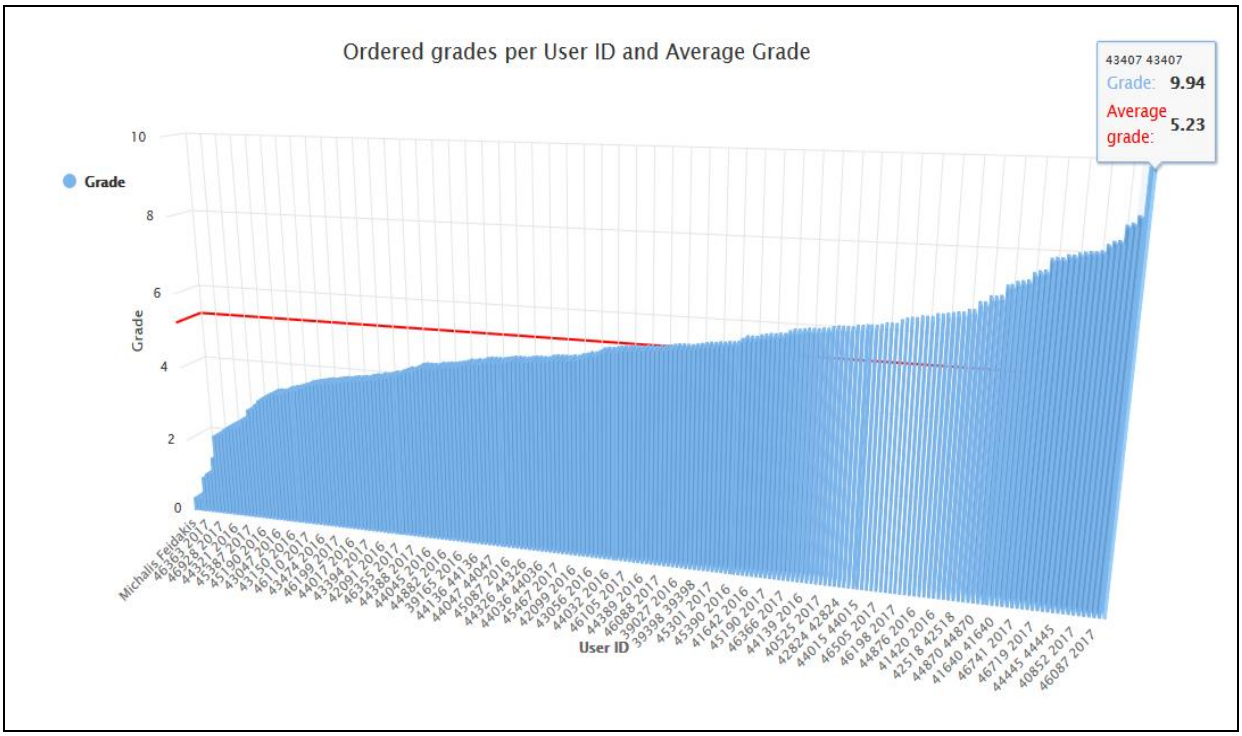
Εδώ εμφανίζονται αναλυτικά για όλα τα επιλεγμένα ακαδημαϊκά έτη συγκριτικά όλες οι βαθμολογίες. Πρόκειται για τρία (3) συνολικά γραφήματα.

Στο πρώτο γράφημα (Εικόνα 6.7, bar chart) για κάθε ακαδημαϊκό έτος εμφανίζονται δύο στήλες, ο αριθμός των εγγεγραμμένων φοιτητών και ο αριθμός των φοιτητών που παρακολούθησαν και βαθμολογήθηκαν (υποσύνολο του πρώτου). Στο δεξί μέρος ο κατακόρυφος άξονας δίνει την κλίμακα, σε πλήθος ατόμων. Διαπιστώνεται ότι οι εγγεγραμμένοι στο μάθημα (γαλάζιες στήλες) ήταν αντίστοιχα περίπου 90 (2015-16), 75 (2016-17), 90 (2017-18) άτομα, αλλά εξ αυτών παρακολούθησαν και βαθμολογήθηκαν περίπου 75, 70 και 75 άτομα, αντίστοιχα (πράσινες στήλες). Στις κατακόρυφες στήλες υπερτίθεται η αριθμητική τιμή της μέσης βαθμολογίας κατ' έτος. Στο αριστερό μέρος ο κατακόρυφος άξονας δίνει τη βαθμολογία σε κλίμακα 0-10. Οι μέσες βαθμολογίες έτους ενώνονται με γραμμή, για οπτικοποίηση της τάσης εξέλιξης εάν υπάρχει. Παρατηρούμε ότι η μέση βαθμολογία κυμάνθηκε περίπου μεταξύ 5,0 και 5,5 στα τρία αυτά έτη.



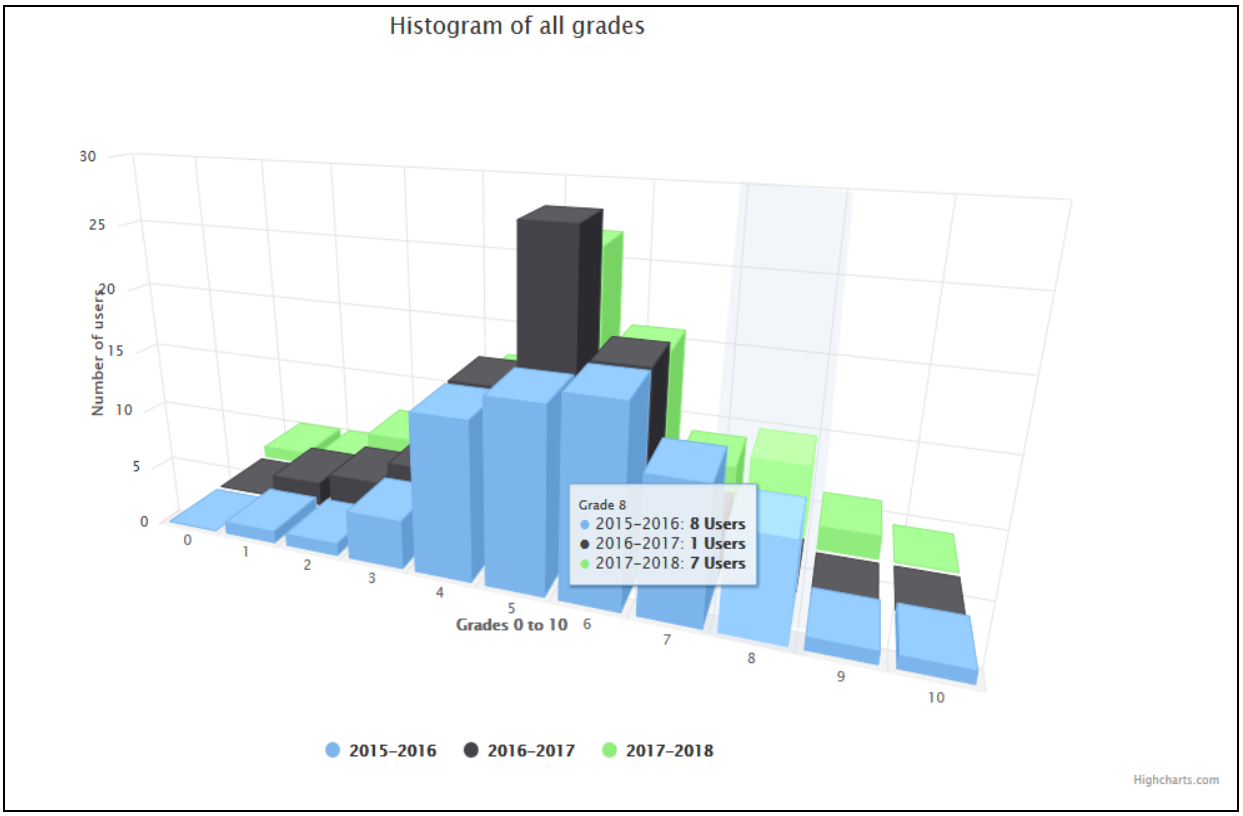
Εικόνα 6.7: Αριθμός εγγεγραμμένων και αριθμός βαθμολογημένων φοιτητών κατ΄ έτος, και μέση βαθμολογία έτους, για όλα τα επιλεγμένα ακαδημαϊκά έτη.

Στο επόμενο γράφημα (Εικόνα 6.8) απεικονίζονται οι ατομικές βαθμολογίες ταξινομημένες από τη μικρότερη προς τη μεγαλύτερη (γαλάζιες στήλες) καθώς και η μέση βαθμολογία (κόκκινη οριζόντια γραμμή) για όλους τους φοιτητές των επιλεγμένων ετών.



Εικόνα 6.8: Επιλεγμένες βαθμολογίες ανά Αναγνωριστικό Χρήστη και Μέση Βαθμολογία για όλα τα ακαδημαϊκά έτη.

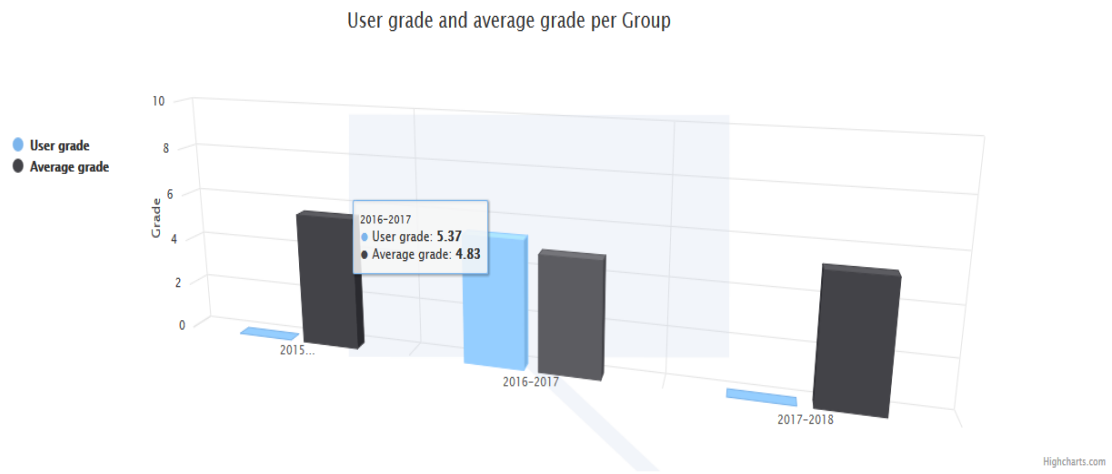
Στο επόμενο γράφημα (Εικόνα 6.9) δίνεται το ιστόγραμμα βαθμολογιών σε 3 διαστάσεις, για όλα τα επιλεγμένα ακαδημαϊκά έτη σε επάλληλα επίπεδα. Οι βαθμολογίες απεικονίζονται σε κλίμακα 0 - 10 (οριζόντιος άξονας) ενώ ο αριθμός των φοιτητών του έτους που επέτυχε την αντίστοιχη βαθμολογία δίνεται στον κατακόρυφο άξονα. Στον 3^{ο} άξονα (βάθος) δίνονται τα επιλεγμένα ακαδημαϊκά έτη. Περνώντας το ποντίκι πάνω από τη στήλη συγκεκριμένης βαθμολογίας, εμφανίζονται σε ορθογώνιο πλαίσιο οι ακριβείς τιμές της βαθμολογίας. Παραδείγματος χάριν, στην Εικόνα 6.8 εμφανίζονται σε πλαίσιο, για τη στήλη του βαθμού οκτώ (8,00), ο ίδιος ο βαθμός, και ο αριθμός φοιτητών που τον επέτυχαν σε κάθε έτος: οκτώ (8) άτομα για το 2015-16, ένα (1) άτομο για το 2016-17 και επτά (7) άτομα για το 2017-18.



Εικόνα 6.9: Ιστόγραμμα των τελικών βαθμολογιών για όλα τα ακαδημαϊκά έτη.

Εξετάζοντας την απεικόνιση από την πλευρά του χρήστη με δικαιώματα φοιτητή, είναι αυτονόητο ότι η απεικόνιση πρέπει να δείχνει όλα τα στοιχεία (και μόνο αυτά) που δίνει η ίδια η Moodle στο ρόλο του φοιτητή – δηλαδή να δει τη δική του βαθμολογία και τον μέσο όρο της τάξης του (Εικόνα 6.10).

Teacher report - 33813 2016



Grade item	Calculated weight	Grade	Range	Percentage	Average	Feedback	Contribution to course total
Ψηφιακή Επεξεργασία Σήματος							
Αυτοαξιολόγηση Section A	-	4.50	0–10	45.00 %	5.84		-
Αυτοαξιολόγηση Section E	-	4.50	0–10	45.00 %	5.37		-
Αυτοαξιολόγηση Section G	-	6.00	0–10	60.00 %	6.59		-
Αυτοαξιολόγηση Section K	-	4.08	0–10	40.83 %	5.25		-
Αυτοαξιολόγηση Section P (Part A)	-	4.67	0–10	46.67 %	5.17		-
Αυτοαξιολόγηση Section P (Part B)	-	6.00	0–10	60.00 %	5.42		-
Αυτοαξιολόγηση Section R	-	7.33	0–10	73.33 %	5.92		-
Αυτοαξιολόγηση Section T	-	5.89	0–10	58.89 %	5.30		-
Course total Include empty grades.	-	**5.37**	**0–100**	**5.37 %**	**5.19**		-

Εικόνα 6.10: Απεικόνιση στοιχείων για χρήστη με δικαιώματα φοιτητή.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Η εκπόνηση μιας μεταπτυχιακής διπλωματικής εργασίας έχει υψηλές απαιτήσεις και θέτει μεγάλες προκλήσεις. Η ανάπτυξη ενός νέου plug-in για μια πλατφόρμα ηλεκτρονικής μάθησης είναι μια απαιτητική προγραμματιστικά εργασία και οι προκλήσεις και οι δυσκολίες που πρόκειται να αντιμετωπίσει ο προγραμματιστής είναι αρκετές – αρκετά και δύσκολα είναι και τα προβλήματα και τα εμπόδια που εμφανίζονται διαρκώς κατά την διαδικασία συγγραφής του κώδικα. Η παρούσα διπλωματική εργασία ξεκίνησε θέτοντας συγκεκριμένους στόχους. Διαπιστώνεται με την ολοκλήρωσή της ότι κατά το μεγαλύτερο ποσοστό τους οι στόχοι αυτοί επετεύχθησαν, αλλά αναπόφευκτα ορισμένοι εξ αυτών παραμένουν ανοικτοί σε περαιτέρω επεξεργασία ώστε να επιτευχθούν στο μέλλον.

Δυσκολίες αντιμετωπίστηκαν με την συμβατότητα των βιβλιοθηκών των προγραμμάτων απεικόνισης, ανάλυσης δεδομένων, ανάκτησης δεδομένων και λογισμικού με την πλατφόρμα της Moodle. Αυτό συμβαίνει διότι η Moodle είναι μια πλατφόρμα ανοικτού κώδικα (open source) με αποτέλεσμα ο προγραμματιστής να πρέπει να έχει τα κατάλληλα συμβατά προγράμματα με την πλατφόρμα για όλες αυτές τις λειτουργίες που αναφέρθηκαν.

Ένας άλλος στόχος που παραμένει ανοικτός είναι η γνήσια 3-D απεικόνιση σμήνους σημείων (3-D scatter plot) που δεν υλοποιήθηκε στο χρονικό πλαίσιο της εργασίας – αν και είναι σήμερα απολύτως εφικτή. Στο πλαίσιο των εκπαιδευτικών δεδομένων τα «σημεία» (observations) αντιστοιχούν συνήθως σε φοιτητές, το μεγάλο πλήθος των οποίων θέτει ειδικές απαιτήσεις για το είδος αυτό της απεικόνισης, ώστε να μην αλληλοεπικαλύπτονται τα σημεία και δίνεται λανθασμένη εικόνα.

Ειδικότερα για το θέμα του 3-D scatter plot, οι δυσκολίες προήλθαν από διαφορετικές αιτίες:

α) Το Moodle χρησιμοποιεί μια σχεσιακή βάση δεδομένων της Oracle. Δυστυχώς διαπιστώθηκε ότι τα ήδη αποθηκευμένα εκπαιδευτικά δεδομένα από τους φοιτητές του Τμήματος είχαν εισαχθεί μάλλον άναρχα στους πίνακες της βάσης αυτής, με αποτέλεσμα να μην μπορούν να ανακτηθούν μαζικά με κατάλληλο προγραμματισμό. Για παράδειγμα, υπήρχαν εγγεγραμμένοι φοιτητές σε διαφορετικά έτη και διαφορετικά μαθήματα, για τους οποίους ως Used ID είχε χρησιμοποιηθεί αλλού το ονοματεπώνυμό τους, αλλού ο αριθμός

μητρώου τους και αλλού ένας αύξων αριθμός εντός του ακαδημαϊκού έτους και μαθήματος. Αυτό απέτρεψε την ενιαία απεικόνιση, διότι θα είχε ως άμεσο αποτέλεσμα να χαθεί η ουσία του απεικονιζόμενου μεγέθους. Δεν θα επιτυγχάνονταν ο βασικός στόχος της Εξόρυξης Δεδομένων, δηλαδή η απεικόνιση ώστε τα δεδομένα να γίνονται γρήγορα και εύκολα κατανοητά από το χρήστη. Αντίθετα, αν επιχειρούνταν ενιαία απεικόνιση σε δεδομένα ίδιας φύσης αλλά διαφορετικής μεθόδου ταυτοποίησης, θα δημιουργούνταν μεγάλη σύγχυση στον χρήστη από τα πυκνά και άναρχα αποτελέσματα.

β) Επίσης ένα άλλο σημαντικό εμπόδιο ήταν το γεγονός ότι σε πολλά από τα ηλεκτρονικά μαθήματα της πλατφόρμας δεν υπήρχαν αποθηκευμένα τα εκπαιδευτικά δεδομένα παλαιοτέρων ετών. Αυτό προγραμματιστικά αποκλείει το μαζικό χειρισμό της βάσης, καθότι η εξαίρεση που θα εισαχθεί ενδεχομένως για ένα μάθημα θα ισχύει για πάντα και δεν μπορεί να έχει όρους υπό συνθήκη. Πρακτικά θα χρειαζόταν να γραφτεί ειδικός κώδικας (ένα ξεχωριστό plug-in) για κάθε ηλεκτρονικό μάθημα.

γ) Τέλος σημαντικό εμπόδιο ήταν και τα εργαλεία απεικόνισης. Πολλά εργαλεία απεικόνισης προσφέρουν στη δωρεάν εκδοχή τους αρκετές λειτουργίες, αλλά συγκεκριμένου στιλ. Η 3-D scatter plot απεικόνιση που επιχειρήθηκε να εφαρμοστεί, ήταν χαρακτηριστικό επί πληρωμή σε όλα τα εργαλεία καθότι αποτελεί ένα εξειδικευμένο χαρακτηριστικό τους. Δεν κατέστη δυνατόν στο χρονικό πλαίσιο της εργασίας να γίνει η διαδικασία προμήθειας ενός τέτοιου εργαλείου λογισμικού από το ίδρυμα.

Αυτά τα τρία κυρίως προβλήματα ανάγκασαν την εργασία να αλλάξει πορεία ως προς τον προγραμματισμό του Plug-in και ως προς τις αρχικές εκτιμήσεις και να επιλεχτούν γραφικές αναπαραστάσεις ουσιαστικά δύο διαστάσεων με προοπτική (θέαση) τριών. Έτσι η εργασία δεν παρέκλινε πολύ από τον αρχικό σχεδιασμό της και η οπτικοποίηση των δεδομένων σε ένα πολύ μεγάλο ποσοστό ικανοποίησε τις αρχικές απαιτήσεις.

Ένα κρίσιμο σημείο που πρέπει να κατανοηθεί από όσους ασχολούνται με την ηλεκτρονική μάθηση και την αποθήκευση ψηφιακών δεδομένων της είναι ότι η οπτικοποίηση των δεδομένων θέτει κάποιες συγκεκριμένες απαιτήσεις στην οργάνωση και λειτουργία της βάσης δεδομένων που υποστηρίζει την πλατφόρμα. Μια από αυτές είναι ότι για να γίνεται η οπτικοποίηση εκπαιδευτικών δεδομένων παλαιότερων ετών με αξιοποιήσιμο συγκριτικά τρόπο, θα πρέπει οι φοιτητές κάθε έτους ή ακαδημαϊκής περιόδου να οργανώνονται σε ένα χωριστό group ανά έτος ή περίοδο, με κατάλληλη ονοματολογία. Αυτή η απαίτηση προκύπτει διότι όταν συγγράφει τον κώδικα της μεθόδου

που θα υλοποιήσει την οπτικοποίηση, ο προγραμματιστής ανάλογα τα διαθέσιμα δεδομένα και τον τρόπο ομαδοποίησής τους καθορίζει και τον κατάλληλο τρόπο οπτικοποίησης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Βιβλία (Books):

[1]. Πετράκη, Δ. (2014). *Η Πλατφόρμα Moodle και η Εφαρμογή της στην Εκπαίδευση*. Πτυχιακή Εργασία, Τεχνολογικό Εκπαιδευτικό Ίδρυμα Ηπείρου (Τ.Ε.Ι Ηπείρου), Τμήμα Μηχανικών Πληροφορικής Τ.Ε. pp. 9-92.

[2]. Abebe, A., Daniels, J., and McKean, J.W. (2001). *Statistics and Data Analysis Statistical Computation Lab (SCL)*, Western Michigan University, USA. pp. 2-112.

[3]. Bacon, J., (2007). *Practical PHP and MySQL-Building Eight Dynamic Web Applications*. Pearson Education, Inc., pp. 5-60/303-376/425-467.

[4]. Kramer, E. (2000). *Οπτικός Οδηγός της HTML 4.* Μετάφραση Μαίρη Γκλαβά, Εκδόσεις Μ. Γκιούρδας. pp. 2-276.

[5]. Lang, C., Siemens, G., Wise, A., Gašević, D. (2017). *Handbook of Learning Analytics*. Solar Edition-Society for Learning Analytics Research, University of Michigan pp. 17-92/163-240/319-355.

[6]. Charles, J. (1985). *Macmillan Dictionary of Data Communications*. The Macmillan Press Ltd. pp. 3-532.

[7]. Nong, Y. (2003). *The Handbook of Data Mining*. Arizona State University - Human Factors and Economics, Lawrence Erlbaum Associates Publishers. pp. 3-65/159-190.

[8]. Valade, J. (2007). *PHP 5 and MySQL for Dummies*. Wiley Publishing Inc. pp. 7-327.

[9]. Welling, L., and Thomson, L. (2008). *Ανάπτυξη Web Εφαρμογών με PHP και MySQL*. Εκδόσεις Μ.Γκιούρδας. pp. 11-78/217-263/323-37.

[10]. Wiley, J. (2015). *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. John Wiley & Sons, Inc. pp. 2-24/30-42/64-110.

Τεχνικά Άρθρα και Αναφορές (Technical Papers & Reports):

[11]. Bikakis, N. (2018). Big Data Visualization Tools. *Encyclopedia of Big Data Technologies*, Springer.

[12]. Charitopoulos, A., Rangoussi, A., Koulouriotis, D. (2017). Educational data mining and data analysis for optimal learning content management – Applied in Moodle for under graduated engineering studies. *IEEE Global Engineering Education Conference (EDUCON 2017)*, Athens, Greece.

[13]. Ferguson, M. (2014). Data Visualization–Flexible Technology for the Agile Enterprise. *Intelligent Business Strategies*, S.A.S.

[14]. Friendly, M. (2006). *A Brief History of Data Visualization: Handbook of Data Visualization*. Springer Handbooks, Comp. Statistics. Springer.

[15]. Pena-Ayala, A. (2013). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 41(A), Part 1, pp1432-1462.

[16]. Pullokkaran. L.J (2013). *Analysis of Data Visualization & Enterprise Data Standardization in Business Intelligence*. Massachusetts Institute of Technology.

[17]. Romero, C. and Ventura, S. (2006). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1), pp.135-146.

[18]. Romero, C., Ventura, S., and Garcıa, E. (2007). Data mining in course management systems: Moodle case study and tutorial. *Computer Sciences & Education,* 51, pp. 368-384.

[19]. David, S. (2011). *Next-Generation Data Visualization*. Netspective Communications LLC.

[20]. Sykamiotis, G., Charitopoulos, A., Rangoussi, M., and Koulouriotis, D. (2017). Extraction and presentation of access and usage data from an e-learning platform (Moodle)-Design and development of a software application. *IEEE Global Engineering Education Conference (EDUCON 2017)*, Athens, Greece.

[21]. Επιτροπή Παιδείας-Τμήμα Σύγχρονων Γλωσσών, Στρασβούργο. Κοινό Ευρωπαϊκό Πλαίσιο αναφοράς για τη γλώσσα: εκμάθηση, διδασκαλία, αξιολόγηση. Council of Europe.

Διαδικτυακές πηγές (Internet Sources):

[22]. https://docs.moodle.org

[23]. https://europa.eu/european-union/index_en

[24]. http://forum-gephi.org/

[25]. https://gephi.org/

[26]. https://plot.ly/

[27]. https://public.tableau.com/en-us/s/

[28]. https://stackoverflow.com/

[29]. https://slideplayer.gr/slide/2853885/

[30]. https://www.datawrapper.de/

[31]. https://www.highcharts.com/

[32]. https://www.macmillan.com

[33]. https://www.mgiurdas.gr/

[34]. https://www.primefaces.org/

[35]. https://www.wikipedia.org/

[36]. http://sci2s.ugr.es/keel/datasets.php

[37]. https://docs.moodle.org/dev/Gradebook_reports

✓ **Κώδικας από το αρχείο teacher_report.php**

```
<?php
/**
 * @package gradereport_teacherreport
 * @copyright 2018
 */
/**
 * Teacher report class with a group user data.
 */
class teacher_report_group_user_data {
   public $id="";
   public $name="";
   public $grade = 0;
   public function __construct($userId, $userName, $userGrade = 0) {
      $this->id = $userId;
      $this->name = $userName;
      $this->grade = $userGrade;
   }
}
/**
 * Teacher report class with group data.
 */
class teacher_report_group_data {
   public $id="";
   public $name="";
```

```
    // an array of teacher_report_group_user_data
    public $usersData=array();
    public $usersCount = 0;
    public $gradedUsersCount = 0;
    public $averageGrade = 0;
    // an array of group's grades distribution
    public $distributionDataArray = array();
    public function __construct($groupId, $groupName) {
        $this->id = $groupId;
        $this->name = $groupName;
        $keys = array(0,1,2,3,4,5,6,7,8,9,10);
        $this->distributionDataArray = array_fill_keys($keys, 0);
    }
}
/**
 * Teacher report class with all groups data.
 */
class teacher_report_all_groups_data {    // an array of teacher_report_group_data
    public $groupsData=array();
    public function __construct() {
    }
    // Return an array with all groups names
    public function getAllGroupsNamesArray() {
        // Initialize array to keep all group names
        $allGroupsNamesArray = array();
        foreach($this->groupsData as $groupData) {
            $groupName = $groupData->name;
            array_push($allGroupsNamesArray, """ . $groupName  . """);
```

```
        }
        return $allGroupsNamesArray;
    }
    // Return an array with all groups number of users
    public function getAllGroupsUsersArray() {
        // Initialize array to keep all groups users
        $countGroupsUsersArray = array();
        foreach($this->groupsData as $groupData) {
            $countGroupUsers = $groupData->usersCount;
            array_push($countGroupsUsersArray, $countGroupUsers);
        }
        return $countGroupsUsersArray;
    }
    // Return an array with all groups number of graded users
    public function getAllGroupsGradedUsersArray() {
        // Initialize array to keep all groups graded users
        $countGroupsGradedUsersArray = array();
        foreach($this->groupsData as $groupData) {
            $countGroupUsers = $groupData->gradedUsersCount;
            array_push($countGroupsGradedUsersArray, $countGroupUsers);
        }
        return $countGroupsGradedUsersArray;
    }
    // Calculate and return an average grade for all groups.
    public function getAllGroupsAverage() {
        $allGroupsAverage = 0;
        $sum = 0;
        $count = 0;
```

```
        foreach($this->groupsData as $groupData) {
            $sum += ($groupData->gradedUsersCount * $groupData->averageGrade);
            $count += ($groupData->gradedUsersCount);
        }
        $allGroupsAverage = $sum / $count;
        if($allGroupsAverage == "") {$allGroupsAverage=0;}
        $allGroupsAverage = round($allGroupsAverage,2);
        return $allGroupsAverage;
    }
    // Return an array with user's grade and average grade per group.
    public function getAllGroupsUserGradesAverage($userId) {
        $allGroupsUserGrades = array();
        $groups = array();
        $groupAverages = array();
        $userGrades = array();
        $count = 0;
        foreach($this->groupsData as $groupData) {
            // $groupUserGrades = array();
            $userGrade = 0;
            if(array_key_exists($userId,$groupData->usersData)) {
                $userGrade = $groupData->usersData[$userId]->grade;
                $count ++;
            }
            array_push($groups, """.$groupData->name."""");
            array_push($groupAverages, $groupData->averageGrade);
            array_push($userGrades, $userGrade);
        }
        array_push($allGroupsUserGrades, $groups);
```

```
        array_push($allGroupsUserGrades, $userGrades);
        array_push($allGroupsUserGrades, $groupAverages);
        if($count>1) {
            print("Found multiple grades for user");
        }
        return $allGroupsUserGrades;
    }
}
```

✓ **Κώδικας από το φάκελο teacher_report.js**

```
/**
 * Average grade and users per group chart.
 *
 * @param {*} allGroupNamesArray
 * @param {*} countGroupUsersArray
 * @param {*} countGroupGradedUsersArray
 * @param {*} averageGroupGradeArray
 */
function groupsAverageHChart (allGroupNamesArray, countGroupUsersArray,
countGroupGradedUsersArray, averageGroupGradeArray) {

    Highcharts.chart('allGroupsHChart', {
    chart: {
        zoomType: 'xy'
    },
    title: {
        text: 'Average grades and Users per Group'
    },
    subtitle: {
        text: ''
    },
    xAxis: [{
        categories: allGroupNamesArray,
        crosshair: true
```

```
}],
yAxis: [{ // Primary yAxis
   labels: {
   format: '{value}',
   style: {
      color: Highcharts.getOptions().colors[1]
   }
   },
   title: {
      text: 'Average grade',
      style: {
         color: Highcharts.getOptions().colors[1]
      }
   },
   min: 0,
   max: 10,
   allowDecimals: false,
}, { // Secondary yAxis
   title: {
   text: 'Users',
   style: {
      color: Highcharts.getOptions().colors[0]
   }
   },
   labels: {
   format: '{value}',
   style: {
      color: Highcharts.getOptions().colors[0]
   }
   },
   opposite: true
}],
tooltip: {
   shared: true
},
series: [{
      name: 'Graded users',
      type: 'column',
      yAxis: 1,
      data: countGroupGradedUsersArray,
```

```
            tooltip: {
            valueSuffix: ''
            },
            color: Highcharts.getOptions().colors[2]
        }, {
            name: 'Users',
            type: 'column',
            yAxis: 1,
            data: countGroupUsersArray,
            tooltip: {
            valueSuffix: ''
            },
            color: Highcharts.getOptions().colors[0]
        }, {
            name: 'Average grade',
            type: 'spline',
            data: averageGroupGradeArray,
            tooltip: {
            valueSuffix: ''
            },
            color: Highcharts.getOptions().colors[1]
        }]
    });
}
/**
 * All selected user's grades 3d perspective chart.
 *
 * @param {*} allUsers
 * @param {*} allUsersAverageGrade
 * @param {*} allUsersGrades
 */
function allUsersHChart3d(allUsers, allUsersAverageGrade, allUsersGrades) {
    Highcharts.chart('allUsersHChart3d', {
        chart: {
            zoomType: 'xy',
            options3d: {
                enabled: true,
                alpha: 15,
                beta: 15,
                depth: 50,
```

```
        viewDistance: 25
      }
    },
    title: {
      text: 'Ordered grades per User ID and Average Grade'
    },
    subtitle: {
      text: ''
    },
    xAxis: [{
      categories: allUsers,
      crosshair: true,
      title: {
        text: 'User ID',
        style: {
          fontWeight: 'bold'
        }
      },
      margin:26,
    }],
    yAxis: [{
      labels: {
        format: '{value}',
        style: { color: Highcharts.getOptions().colors[1] }
      },
      title: {
        text: 'Grade',
        style: { color: Highcharts.getOptions().colors[1] }
      },
      plotLines:[{
        value: allUsersAverageGrade[0],
              width: 2,
        color: 'red'
      }],
      min: 0,
      max: 10,
      allowDecimals: false,
    }],
    tooltip: {
      shared: true,
```

```
        useHTML: true,
        headerFormat: '<small>{point.key}</small><table>',
        pointFormat: '<tr><td style="color: {series.color}"><small>{series.name}:
</small></td>' +
                    '<td style="text-align: right"><small><b>{point.y}</b></small></td></tr>' +
                    '<tr><td style="color: red"><small>Average grade: </small></td>' +
                    '<td style="text-align: right"><small><b>'+ allUsersAverageGrade[0]
+'</b></small></td></tr>',
        footerFormat: '</table>',
        valueDecimals: 2
      },
      legend: {
        layout: 'vertical',
        align: 'left',
        x: 100,
        verticalAlign: 'top',
        y: 100,
        floating: true,
        backgroundColor: (Highcharts.theme && Highcharts.theme.legendBackgroundColor)
|| '#FFFFFF'
      },
      plotOptions: {
        column: {
          pointPadding: 0.2,
          borderWidth: 0
        }
      },
      series: [{
        name: 'Grade',
        type: 'column',
        data: allUsersGrades,
        tooltip: {
          valueSuffix: ''
        }
      }]
    });
}

/**
 * Distribution per group 3d perspective chart.
```

```
 *
 * @param {*} distGroupsNamesArray
 * @param {*} distGroupsDataArray
 */
function distributionPerGroupHChart3d (distGroupsNamesArray, distGroupsDataArray) {
   var seriesArray =[];
   for ( i=0; i<distGroupsNamesArray.length; i++) {
      seriesN =
         {
         name: distGroupsNamesArray[i],
         data: distGroupsDataArray[i],
         stack: i,
         tooltip: {
            enabled: true,
            valueSuffix: ' Users'
         }
      };
      seriesArray.push(seriesN);
   }
   var hchart = Highcharts.chart('distributionPerGroupHChart3d', {
      chart: {
         type: 'column',
         margin: 75,
         options3d: {
          enabled: true,
            alpha: 20,
            beta: 20,
            depth: 40,
            viewDistance: 25,
            frame: {
               bottom: {
                  size: 1,
                  color: 'rgba(0,0,0,0.05)'
               }
            }
         }
      },
      title: {
         text: 'Histogram of all grades'
      },
```

```
plotOptions: {
    column: {
        depth: 100,
        stacking: true,
        grouping: false,
        groupPadding: 0,
        groupZPadding: 10
    }
},
xAxis: [{
    categories: [0,1,2,3,4,5,6,7,8,9,10],
    crosshair: true,
    title: {
        text: 'Grades 0 to 10',
        style: {
            fontWeight: 'bold'
        }
    },
}],
yAxis: [{ // Primary yAxis
    labels: {
        format: '{value}',
        style: {
            color: Highcharts.getOptions().colors[1]
        }
    },
    title: {
        text: 'Number of users',
        style: {
            color: Highcharts.getOptions().colors[1]
        }
    },
    min: 0,
    allowDecimals: false,
}],
tooltip: {
    shared: true,
    headerFormat: '<span style="font-size: 10px">Grade {point.key}</span><br/>'
},
series: seriesArray
```

```
    });
}
/**
 * User grades and average grades per group.
 * Column 3d chart.
 *
 * @param {*} userId
 * @param {*} allGroupNamesArray
 * @param {*} userGradeArray
 * @param {*} averageGroupGradeArray
 */
function userHChart3d(userId, allGroupNamesArray, userGradeArray,
averageGroupGradeArray) {
    container = 'userHChart3d'+ userId;
    Highcharts.chart(container, {
    chart: {
        type: 'column',
        options3d: {
            enabled: true,
            alpha: 15,
            beta: 15,
            depth: 50,
            viewDistance: 25
        }
    },
    title: {
        text: 'User grade and average grade per Group'
    },
    subtitle: {
        text: ''
    },
    xAxis: [{
        categories: allGroupNamesArray,
        crosshair: true
    }],
    yAxis: [{ // Primary yAxis
        labels: {
            format: '{value}',
            style: {
                color: Highcharts.getOptions().colors[1]
```

```
            }
        },
        title: {
            text: 'Grade',
            style: {
                color: Highcharts.getOptions().colors[1]
            }
        },
        min: 0,
        max: 10,
        allowDecimals: false,
    }],
    tooltip: {
        shared: true
    },
    legend: {
        layout: 'vertical',
        align: 'left',
        x: 100,
        verticalAlign: 'top',
        y: 100,
        floating: true,
        backgroundColor: (Highcharts.theme && Highcharts.theme.legendBackgroundColor) ||
'#FFFFFF'
    },
    series: [{
            name: 'User grade',
            data: userGradeArray,
            tooltip: {
                valueSuffix: ''
            }
        }, {
            name: 'Average grade',
            data: averageGroupGradeArray,
            tooltip: {
                valueSuffix: ''
            }
        }]
    });
```

Access protocol: The set of rules that workstations use to avoid collisions when sending information over shared network media. Also known as the media access control protocol.

Access server: A computer that provides access for remote users who dial in to the system and access network resources as though their computers were directly attached to the network.

Access time: The period of time that elapses between a request for information from disk or memory and the arrival of that information at the requesting device. Memory-access time refers to the time it takes to transfer a character between memory and the processor. Disk-access time refers to the time it takes to place the read/write heads over the requested data. RAM may have an access time of 80 nanoseconds or less, while hard-disk access time could be 10 milliseconds or less.

Account policy: On networks and multiuser operating systems, the set of rules that defines whether a new user is permitted access to the system and whether an existing user is granted additional rights or expanded access to other system resources. Account policy also specifies the minimum length of passwords, the frequency with which passwords must be changed, and whether users can recycle old passwords and use them again.

Active Server Pages: Abbreviated ASP. In Microsoft Internet Information Server, a script interpreter and execution environment that supports VBScript and Java- Script and is compatible with other scripting languages such as Perl, REXX, Tcl, and Python through add-ins from third-party developers. ASP allows you to combine HTML, scripts, and ActiveX components on the same Web server; all the code runs on the server and presents the results of this dynamic process to the client browser as a standard HTML page.

Address: 1. The precise location in memory or on disk where a piece of information is stored. Each byte in memory and each sector on a disk have its own unique address. **2.** The unique identifier for a specific node on a network. An address may be a physical address specified by switches or jumpers on the network interface card hardware, or it can be a logical address established by the network operating system. **3.** To reference or manage a storage location. **4.** In UNIX, an IP address as specified in the /etc/hosts file. **5.** Information used by a network or the Internet to specify a specific location in the form username@hostname; username is your user name, logon name, or account name or number, and hostname is the name of the Internet Service Provider (ISP) or computer system you use. The hostname may consist of several parts, each separated from the next by a period.

Advanced Communications Service: Abbreviated ACS. A large data-communications network established by AT&T.

Advanced Configuration and Power Interface: Abbreviated ACPI. An interface specification developed by Intel, Microsoft, and Toshiba for controlling power use on the PC and all other devices attached to the system. A BIOS-level hardware specification, ACPI depends on specific hardware that allows the operating system to direct power management and system configuration.

Advanced Data Communications Control Procedures: Abbreviated ADCCP. A bit-oriented, link-layer, ANSI-standard communications protocol.

Advanced Peer-to-Peer Networking: Abbreviated APPN. IBM's SNA (Systems Network Architecture) protocol, based on APPC (Advanced Program-to-Program Communications). APPN allows nodes on the network to interact without a mainframe host computer and implements dynamic network directories and dynamic routing in an SNA network. APPN can run over a variety of network media, including Ethernet, token ring, FDDI, ISDN, X.25, SDLC, and higher-speed links such as B-ISDN or ATM.

Angle brackets: The less-than (<) and greater-than (>) symbols used to identify a tag in an HTML document. Also used to identify the return address in an e-mail message header.

Anonymous FTP: A method used to access an Internet computer that does not require you to have an account on the target computer system. Simply log on to the Internet computer with the user name *anonymous,* and use your e-mail address as your password. This access method was originally provided as a courtesy so that system administrators could see who had logged on to their systems, but now it is often required to gain access to an Internet computer that has FTP service.

Anonymous server: A special Usenet service that removes from a Usenet post all header information that could identify the original sender and then forwards the message to its final destination. If you use an anonymous server, be sure to remove your signature from the end of the message; not all anonymous servers look for and then strip a signature. Also known as an anonymous remailer.

Apache HTTP Server: A freeware Web server, supported by the Unix community, in use on almost half of the Web sites on the Internet. So called because the original university- lab software was patched with new features and fixes until it became known as "a patchy server."

Application: Abbreviated app, or if the application is a small one, it is referred to as an applet. A computer program designed to perform a specific task, such as accounting, scientific analysis, word processing, or desktop publishing. In general, applications can be distinguished from system software, system utilities, and computer language compilers, and they can be categorized as either stand-alone or network applications. Stand-alone applications run from the hard disk in an independent computer, so only one user at a time can access the application. Network applications run on networked computers and can be shared by many users. Advanced applications such as groupware and e-mail allow communications between network users.

Application layer: The seventh, or highest, layer in the OSI Reference Model for computer-to-computer communications. This layer uses services provided by the lower layers but is completely insulated from the details of the network hardware. It describes how applications interact with the network operating system, including database management, electronic mail, and terminal emulation programs.

Application object: In Novell Directory Services (NDS), a leaf object that represents a network application in a NetWare Directory tree.

Application server: A special-purpose file server that is optimized for a specific task, such as communications or a database application, and that uses higher-end hardware than a typical file server.

Associated Accredited Systems Engineer: Abbreviated AASE. A certification from Compaq designed to evaluate and recognize basic knowledge of PC architecture and operations. An AASE may choose to specialize in Microsoft Windows 2000 or Novell NetWare operation.

Asymmetric Digital Subscriber Line: Abbreviated ADSL. A high-speed data transmission technology originally developed by Bellcore and now standardized by ANSI as T1.413. ADSL delivers high bandwidth over existing twisted-pair copper telephone lines. Also called Asymmetric Digital Subscriber Loop. ADSL supports speeds in the range of 1.5 to 9Mbps in the downstream direction (from the network to the subscriber) and supports upstream speeds in the range of 16 Kbps to 640 Kbps; hence, the term *asymmetric*.

Asynchronous communications server: A LAN server that allows a network user to dial out of the network into the public switched telephone system or to access leased lines for asynchronous communications. Asynchronous communications servers may also be called dial-in/dial out servers or modem servers.

AT command set: A set of standard instructions used to activate features on a modem. Originally developed by Hayes Microcomputer Products, the AT command set is now used by almost all modem manufacturers.

ATM Adaptation Layer: Abbreviated AAL. A service-dependent layer in Asynchronous Transfer Mode (ATM) that provides the protocol translation between ATM and the other communications services involved in a transmission. AAL has several service types and classes of operation to handle different kinds of traffic, depending on how data is transmitted, the bandwidth required, and the types of connection involved.

Authorization: The provision of rights or permissions based on identity. Authorization and authentication go hand in hand in networking; your access to services is based on your identity, and the authentication processes confirm that you are who you say you are.

Auto-answer: A feature of a modem that allows it to answer incoming calls automatically.

Automatic Client Upgrade: A mechanism used to upgrade Novell client software during the logon process by executing four separate programs called by the logon script. Automatic Client Upgrade can be very useful when all client workstations use standard configurations.

Automatic forwarding: A feature of many e-mail programs that automatically retransmits incoming messages to another e-mail address.

Automatic rollback: In a Novell Net- Ware network, a feature of the Transaction Tracking System (TTS) that abandons the current transaction and returns a database to its original condition if the network fails in

the middle of a transaction. Automatic rollback prevents the database from being corrupted by information from incomplete transactions.

Autonomous System Border router: Abbreviated ASBR. In an internetwork that uses link state routing protocols such as Open Shortest Path First (OSPF) protocols, a router that has at least one connection to a router in an external network.

Back-end processor: A secondary processor that performs one specialized task very effectively, freeing the main processor for other, more important work.

Back-end system: The server part of a client/server system that runs on one or more file servers and provides services to the front-end applications running on networked workstations. The back-end system accepts query requests sent from a frontend application, processes those requests, and returns the results to the workstation. Back-end systems may be PC-based servers, super servers, midrange systems, or mainframes.

Background: 1. On a computer screen, the color on which characters are displayed; for example, white characters may appear on a blue background. **2.** In an operating system, a process that runs in the background generally runs at lower level of priority than a foreground task and does not accept input from the user. Only multitasking operating systems support true background and foreground processing, but some applications can mimic it. For example, many word processors can print a document while still accepting input from the keyboard. In older systems, a process spends its entire existence in either the background or the foreground; in newer systems, you can change the processing environment and move foreground programs into the background, and vice versa.

Backing out: The process of abandoning the current transaction and returning a database to its original condition if the network fails during the transaction. This process prevents the database from being corrupted by information from the incomplete transaction.

Backup: An up-to-date copy of all your files. You make a backup for several reasons:

- Insurance against possible hard-disk or file-server failure. Hard disks often fail completely, taking all your work with them. If this failure occurs, you can reload your files and directories from the backup copy. A backup is your insurance against disk failure affecting the thousands or possibly tens of thousands of files you might have on your file server.
- Protection against accidental deletion of files or directories. Again, if you mistakenly delete a file or directory, you can retrieve a copy from your last backup.
- Protection against the new version of software you are about to install not working to your expectations; make a backup before installing new software.
- As an archive at the end of a project, when a person leaves your company, or at the end of a financial period such as year-end close. Your decision when or how often to make a backup depends on how frequently important data on your system changes. If you rely on certain files always being available on your system, it is crucial that you make regular, consistent backups. Here are some backup tips:
- Keep multiple copies; redundancy should be a part of your backup plan.
- Test your backups to make sure they are what you think they are before you bring the server into service, and make sure you can reload the information you need.
- Store your backups in a secure off-site location; do not leave them right next to the computer (if the computer is damaged by an accident, the backup may be damaged as well).
- Replace your backup media on a regular basis.

Backup browser: In Microsoft Windows 2000 Server, a computer that maintains a list of computers and services available on the network. This list is provided by the master browser and distributed to a workgroup or a domain by the backup browser.

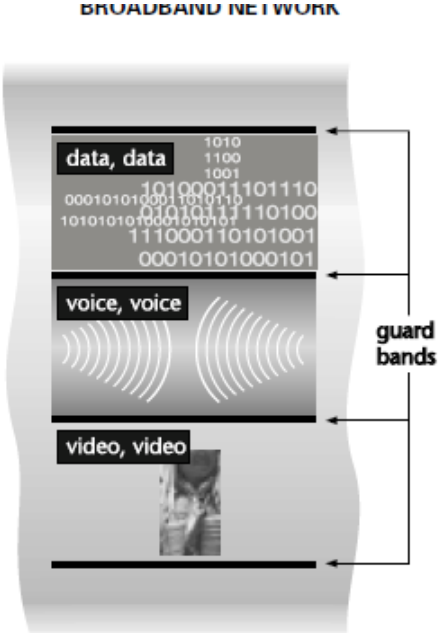
Backup domain controller: In Microsoft Windows NT, a server containing accurate replications of the security and user databases. The backup domain controller receives a copy of the domain's directory database, containing all the account and security information for the domain, from the primary domain controller. This copy is periodically synchronized with the original master database. A domain can contain several backup domain controllers.

Beta software: Software that has been released to a cross-section of typical users for testing before the commercial release of the package.

Bit-oriented protocol: A communications protocol in which data is transmitted as a stream of bits rather than as a stream of bytes. A bit-oriented protocol uses specific sequences of bits as control codes, unlike a

byte-oriented protocol, which uses reserved characters. HDLC (High-level Data Link Control) and IBM's SDLC (Synchronous Data Link Control) are both bit-oriented protocols.

Broadband network: A technique for transmitting a large amount of information, including voice, data, and video, over long distances using the same communications channel. Sometimes called wideband transmission, it is based on the same technology used by cable television. The transmission capacity is divided into several distinct channels that can be used concurrently by different networks, normally by frequency-division multiplexing (FDM). The individual channels are protected from each other by guard channels of unused frequencies. A broadband network can operate at speeds of up to 20Mbps.



Browser: 1. An application program used to explore Internet resources. A browser lets you wander from Web site to Web site without concern for the technical details of the links between them or the specific methods used to access them and presents the information— text, graphics, sound, or video—as a document on the screen. **2.** A small application used to scan a database or a list of files. **3.** In Windows NT networking, a mechanism used as a name service.

Buffer: An area of memory set aside for temporary storage of data. Often, the data remains in the buffer until some external event finishes. A buffer can compensate for the differences in transmission or processing speed between two devices or between a computer and a peripheral device, such as a printer. Buffers are implemented in a variety of ways, including first-in-first-out (FIFO) used for pipes and last-in-last-out used for stacks and circular buffers such as event logs.

Certified Application Developer for Developer/2000: A certification from Oracle consisting of a set of exams covering Structured Query Language, the creation of procedures using Oracle Procedure Builder, using Developer/2000, and managing the user interface.

Certified Computing Professional Abbreviated CCP. A certification from the Institute for Certification of Computing Professionals designed for experienced professionals with more than four years experience in a wide variety of computing and related tasks.

Certified Database Administrator: 1. Abbreviated CDA. A certification from Oracle that covers knowledge of Structured Query Language, administration of Oracle products, along with backup and recovery, and system performance tuning. **2.** Abbreviated CDA. A certification from Sybase that covers designing, building, and supporting Sybase SQL Server databases.

Certified Information System Auditor: Abbreviated CISA. A certification from the Information Systems Audit and Control Association (ISACA) that covers ethics, security, system organization and management, and system development, acquisition, and maintenance.

Character: A symbol that corresponds to a key on the keyboard. A character can be a letter, a number, punctuation, or a special symbol and is usually stored as a single byte. A collection of related characters is known as a *character set*, and the most common character set on PC systems is the American Standard Code for Information Interchange (ASCII). Some larger IBM systems still use Extended Binary

Coded Decimal Interchange Code (EBCDIC). In an attempt to rationalize the many international character sets in use these days, some systems use more than one byte to store a character.

Client application: In OLE, the application that starts a server application to manipulate linked or embedded information.

Client pull: A mechanism used on the Internet whereby a client application, usually a Web browser, initiates a request for services from a Web site.

Cluster controller: An IBM or IBM-compatible device located between a group of 3270 terminals and the mainframe computer. The cluster controller communicates between the computer and the terminals using SDLC (Synchronous Data Link Control) or a bisynchronous communications protocol.

Clustering: A fault-tolerant technology designed to keep server availability at a very high level. Clustering group's servers and other network resources into a single system; if one of the servers in the cluster fails, the other servers can take over the workload. Clustering software also adds a load-balancing feature to make sure that processing is distributed in such as way as to optimize system throughput.

Collision: In networking or communications, an attempt by two nodes to send a message on the same channel at exactly the same moment.

Command line: Any interface between the user and the command processor that allows you to enter commands from the keyboard for execution by the operating system.

Command prompt: A symbol (character or group of characters) on the screen that lets you know that the operating system is available and ready to receive input.

Commercial Internet Exchange: Abbreviated CIX, pronounced "kicks." A connection point between ISPs. A location where top-tier ISPs maintain the routers used to route packets between their respective network segments.

Common Internet File System: Abbreviated CIFS. A file system supported by Microsoft, DEC, Data General, SCO, and others, which allows users and organizations to run file systems over the Internet. CIFS is an extension to Microsoft's Server Message Blocks (SMB) file-sharing protocol and allows users to share files over the Internet in the same way that they share files using networking services on Windows clients.

Compiler: A program that converts a set of program language source code statements into a machine-readable form suitable for execution by a computer. Most compilers do much more than this, however; they translate the entire program into machine language, while at the same time, and they check your source code syntax for errors and then post error messages or warnings as appropriate.

Compression Control Protocol: Abbreviated CCP. A protocol used with Point to- Point Protocol (PPP) to configure, enable, and disable data compression algorithms at both ends of the point-to-point connection. CCP can support different compression algorithms in each direction of the connection.

Configuration: The process of establishing your own preferred setup for an application, expansion board, computer system, or network. Most current software can establish a configuration for you automatically, although you may need to adjust that configuration to get the best results.

Connectivity: The degree to which any given computer or application can cooperate with other network components purchased from other vendors, in a network environment in which resources are shared.

Control code: A sequence of one or more characters used for hardware control; also known as setup strings or escape sequences. Control codes are used with printers, modems, and displays. Printer control codes often begin with an escape character, followed by one or more characters that the printer interprets as commands it must perform rather than as text it must print.

Control Panel: In Microsoft Windows, a special system folder that contains applets used to look at or change configuration information. Each applet manages a single task such as adding or removing a program from your system, setting up a connection to the Internet, or changing display settings

Cookie: 1. A block of data sent from a server to a client in response to a request by the client. **2.** On the World Wide Web, a block of data stored by the server on the system running the browser or client software, which can be retrieved by the server during a future session. A cookie contains information that can identify the user for administrative reasons or to prepare a custom Web page.

Cursor: A special character displayed on a monitor to indicate where the next character will appear when it is typed. In text or character mode, the cursor is usually a blinking rectangle or underline. In a graphical

user interface, the mouse cursor can take many shapes, depending on the current operation and its screen location.

Data: Information in a form suitable for processing by a computer, such as the digital representation of text, numbers, graphic images, or sounds. Strictly speaking, data is the plural of the Latin word *datum*, meaning an item of information; but the term is commonly used in both plural and singular constructions.

Database: A collection of related objects, including tables, forms, reports, queries, and scripts, created and organized by a database management system (DBMS). A database can contain information of almost any type, such as a list of magazine subscribers, personal data on the space shuttle astronauts, or a collection of graphical images and video clips.

Database management system: Abbreviated DBMS. Software that controls the data in a database, including overall organization, storage, retrieval, security, and data integrity. A DBMS can also format reports for printed output and can import and export data from other applications using standard file formats. A data-manipulation language is usually provided to support database queries.

Database model: The method used by a database management system (DBMS) to organize the structure of the database. The most common database model is the relational database.

Database server: Any database application that follows the client/server architecture model, which divides the application into two parts: a front-end running on the user's workstation and a back-end running on a server or host computer. The front-end interacts with the user and collects and displays the data. The back-end performs all the computer-intensive tasks, including data analysis, storage, and manipulation.

Datagram Delivery Protocol: Abbreviated DDP. A routing protocol developed by Apple Computer as a part of its AppleTalk network.

Data-link layer: The second of seven layers of the OSI Reference Model for computer- to-computer communications. The data link layer validates the integrity of the flow of data from one node to another by synchronizing blocks of data and controlling the flow of data. The Institute of Electrical and Electronic Engineers (IEEE) has divided the data-link layer into two other layers—the logical link control (LLC) layer sits above the media access control (MAC) layer.

Data mining: The process of displaying historical commercial data in a multidimensional form so that previously hidden relationships are exposed through the use of advanced statistical tools, making them easier to group and summarize.

Data warehousing: A method of storing very large amounts of data, usually historical transaction processing data, for later analysis and reporting. The data warehouse is accessed by software capable of extracting trends from the raw data and creating comparative reports.

DB connector: Any of several types of cable connectors used for parallel or serial cables. The number following the letters DB (for data bus) indicates the number of pins that the connector usually has; a DB-25 connector can have a maximum of 25 pins, and a DB-9 connector can have as many as 9. In practice, not all the pins (and not all the lines in the cable) may be present in the larger connectors. If your situation demands that all the lines be present make sure you buy the right cable. Common DB connectors include the following:

- **DB-9** Defined by the RS-449 standard as well as the ISO (International Organization for Standardization).
- **DB-25** A standard connector used with RS-232-C wiring, with 25 pins (13 on the top row and 12 on the bottom).
- **DB-37** Defined as the RS-449 primary channel connector.
- **DB-15, DB-19, and DB-50** connectors are also available. The accompanying illustration shows a male and female DB-25 connector.

Decryption: The process of converting encrypted data back into its original form.

Default server: In Novell NetWare, the server that responds to the Get Nearest Server request made as a user starts the logon process. Novell Directory Services has replaced the need for the default server destination with the default context.

Demand paging: A common form of virtual memory management in which pages of information are read into memory from disk only when required by the program.

DeskSet: A collection of graphical desktop applications bundled with Sun Microsystems Solaris. DeskSet includes a file manager with options for copying, moving, renaming, and deleting files, a terminal emulator, text editor, calculator, clock, and calendar, as well as special programs and utilities.

Desktop management: The process of managing desktop workstation hardware and software components automatically, often from a central location.

Directory Access Protocol: Abbreviated DAP. A mail standard used to access white page directories containing names, addresses, e-mail addresses, and telephone numbers. Because of its complexity, DAP has been largely replaced by Lightweight Directory Access Protocol (LDAP).

DNS name server: A server containing information that is part of the Domain Name Service (DNS) distributed database, which makes computer names available to client programs querying for name resolution on the Internet.

DNS Server log: In Microsoft Windows 2000 Server, a special log that records any events associated with running the Domain Name Service (DNS) Server.

Domain: A description of a single computer, a whole department, or a complete site, used for naming and administrative purposes. Top-level domains must be registered to receive mail from outside the organization; local domains have meaning only inside their own enterprise. Depending on the context, the term can have several slightly different meanings:

- On the Internet, a domain is part of the Domain Name Service (DNS).
- In Novell NetWare, a domain is a special area of memory where a new NetWare Loadable Module (NLM) can be tested without the risk of corrupting the operating system memory.
- In IBM's Systems Network Architecture (SNA), a domain represents all the terminals and other network resources controlled by a single processor or processor group.
- In Microsoft Windows NT, a group of computers, users, and network peripherals managed with a single set of account descriptions and security policies. A user can log on to the local computer and be authenticated to access just that one system, or a user can log on to a domain and be authenticated to access other servers within that domain.
- In Lotus Notes, a domain is one or more Notes servers that share the same Public Name and Address Book database. This database contains information about the users within the domain, including their email addresses and other information.

Domain name: In the Domain Name Service (DNS), an easy-to-remember name that identifies a specific Internet host, as opposed to the hard-to-remember numeric IP address.

Domain Name Service: Abbreviated DNS sometimes referred to as Domain Naming System. A distributed addressing system that resolves the domain name into the numeric IP address. DNS lets you use the Internet without having to remember long lists of cryptic numbers. The most common high-level domains on the Internet include:

- .com A commercial organization.
- .edu An educational establishment such as a university.
- .gov A branch of the U.S. government.
- .int An international treaty organization.
- .mil A branch of the U.S. military.
- .net A network provider.
- .org A nonprofit organization Most countries also have unique domains named after their international abbreviation— for example: .uk for the United Kingdom and .ca for Canada.

Dynamic Data Exchange: Abbreviated DDE. A technique used for application-to application communications, available in several operating systems, including Microsoft Windows, Macintosh, and OS/2. When two or more programs that support DDE are running at the same time, they can exchange data and commands, by means of conversations. A DDE conversation is a two-way connection between two applications, used to transmit data by each program alternately. DDE is used for low-level communications that do not need user intervention. For example, a communications program might feed stock market information into a spreadsheet program, where that data can be displayed in a meaningful way and recalculated automatically as it changes. DDE has largely been superseded by Object Linking and Embedding (OLE).

Dynamic DNS: Abbreviated DDNS. In Microsoft Windows 2000 Server, a mechanism that allows Dynamic Host Configuration Protocol (DHCP) and Windows 2000 clients to update Domain Name Service (DNS) records dynamically, rather than by the traditional method of manually adding the new records to static DNS zone files.

Emulator: A device built to work exactly like another device—hardware, software, or a combination of both. For example, a terminal emulation program lets a PC pretend to be a terminal attached to a mainframe

computer or to an online service by providing the control codes that the remote system expects to receive. In printers, some brands emulate popular models such as Hewlett-Packard's LaserJet line.

Encryption: The process of encoding information in an attempt to make it secure from unauthorized access, particularly during transmission. The reverse of this process is known as decryption. Two main encryption schemes are in common use:

- **Private (Symmetrical) Key:** An encryption algorithm based on a private encryption key known to both the sender and the recipient of the information. The encrypted message is unreadable and can be transmitted over no secure systems.
- **Public (Asymmetrical) Key:** An encryption scheme based on using the two halves of a long bit sequence as encryption keys. Either half of the bit sequence can be used to encrypt the data, but the other half is required to decrypt the data.

End user: Often refers to people who us an application to produce their own results on their own computer or workstation. During the mainframe computer era, end users were people who received output from the computer and used that output in their work. They rarely, if ever, actually saw the computer, much less learned to use it themselves. Today, end users often write macros to automate complex or repetitive tasks and sometimes write procedures using command languages.

File format: A file structure that defines the way information is stored in the file and how the file appears on the screen or on the printer. The simplest file format is a plain ASCII file. Some of the more complex formats are DCA (Document Content Architecture) and RTF (Rich Text Format), which include control information for use by a printer; TIFF (Tagged Image File Format) and EPS (Encapsulated PostScript), which hold graphics information; and DBF (Xbase database file) and DB (Paradox file), which are database formats. Word processing programs, such as Microsoft Word, also create files in special formats.

Filename extension: In the MS-DOS file allocation table (FAT) file system, an optional three-character suffix added to the end of a filename and separated from the name by a period.

File permissions: A set of permissions associated with a file (or a directory) that specifies who can access the file and in what way. There are three basic permissions:

- Read permission lets you read files.
- Write permission lets you write (or overwrite) files.
- Execute permission lets you execute files. Additional file permissions vary according to the operating system in use and the security system in place. For example, Novell NetWare has the following file permissions: Access Control, Create, Erase, File Scan, Modify, Read, Supervisor, and Write. Microsoft Windows 2000 has Add and Read, Change Permissions, Delete, Full Control, No Access, and Take Ownership.

File sharing: The sharing of files over a network or between several applications running on the same workstation. Shared files can be read, reviewed, and updated by more than one individual. Access to the file or files is often regulated by password protection, account or security clearance, or file locking to prevent simultaneous changes by multiple users.

Forwarding: The process of passing data on to an intermediate or final destination. Forwarding takes place in network bridges, routers, and gateways.

Front-end application: An application running on a networked workstation that works in conjunction with a back-end system running on the server. Examples are email and database programs.

Full-page display: Any monitor capable of displaying a whole page of text. Full page displays are useful for graphical art and desktop publishing applications, as well as medical applications.

Gopher: A client/server application that presents Internet text resources as a series of menus, shielding the user from the underlying mechanical details of IP addresses and different access methods. Gopher menus may contain documents you can view or download, searches you can perform, or additional menu selections. When you choose one of these items, Gopher does whatever is necessary to obtain the resource you requested, either by downloading a document or by jumping to the selected Gopher server and presenting its top level menu. Gopher clients are available for most popular operating systems, including the Macintosh, MS-DOS, Windows, and Unix.

Graphical user interface: Abbreviated GUI, pronounced "gooey." A graphics based user interface that allows users to select files, programs, and commands by pointing to pictorial representations on the screen rather than by typing long, complex commands from a command prompt. Applications execute in

windows, using a consistent set of drop-down menus, dialog boxes, and other graphical elements, such as scroll bars and icons. This consistency among interface elements is a major benefit

for the user, because as soon as you learn how to use the interface in one program, you can use it in all other programs running in the same environment.

Hacker: In the programming community, where the term originated, this term describes a person who pursues knowledge of computer systems for its own sake—someone willing to "hack through" the steps of putting together a working program.

High-level language: Any machine independent programming language that uses English-like syntax in which each statement corresponds to many assembly language instructions. High-level languages free programmers from dealing with the underlying machine architecture and allow them to concentrate on the logic of the problem at hand.

Hit: On the World Wide Web, a request from a browser for a file on the server. A hit on a Web page occurs whenever any file is accessed, whether it is a text document, a graphic, a script, or an audio or video clip. If you access three files on a Web page, you generate three hits, so a hit is a poor measure of the number of people visiting a Web site, as it simply reflects the number of file requests made.

Home page: On the World Wide Web, an initial starting page. A home page may be prepared by an individual or by a corporation and is a convenient jumping-off point to other Web pages or Internet resources.

Hypertext A method of presenting information so that it can be viewed by the user in a non sequential way, regardless of how the topics were originally organized.

Hyper Text Markup Language: Abbreviated HTML. A standard document formatting language used to create Web pages and other hypertext documents. HTML is a subset of Standardized General Markup Language (SGML).

Hypertext Transfer Protocol: Abbreviated HTTP. The command and control protocol used to manage communications between a Web browser and a Web server.

Integrated software: Application software that combines the functions of several major applications, such as a spreadsheet, a database, a word processor, and a communications program, into a single package. Microsoft Works is an example of integrated software.

Interface: The point at which a connection is made between two hardware devices, between a user and a program or operating system, or between two applications.

Intelligent terminal: A terminal connected to a large computer, often a mainframe that has some level of local computing power and can perform certain operations independently from the remote computer, but does not usually have any local disk-storage capacity.

Internet: The world's largest computer network, consisting of millions of computers supporting tens of millions of users in hundreds of countries. The Internet is growing at such a phenomenal rate that any size estimates are quickly out of date.

Internet address: A location on the Internet. An Internet address takes the form someone@abc.def.xyz, in which someone is a user's name or part of a user's name, @abc is the network computer of the user, and def is the name of the host organization. The last three letters denote the kind of institution the user belongs to:

- edu for educational
- com for commercial
- gov for government
- mil for the military
- org for non-profit organizations
- net for Internet administrative organizations

Internet Explorer: Abbreviated IE. In Microsoft Windows, a Web browser used to display Internet resources.

Internet file types: The Internet offers many opportunities for downloading files from a huge number of Internet hosts. These files may have been generated on different computer systems, so before you spend time downloading a file, it is important to understand the type of file you are dealing with. Table I.1 lists many of the common file types you may encounter.

Intranet: A private corporate network that uses Internet software and TCP/IP networking protocol standards.

JavaScript: A simple scripting language created by Netscape Communications and Sun Microsystems that allows developers to add a specific capability to a Web page. JavaScript is relatively easy to write when

compared with the Java programming language, but it is slower in execution and has far fewer application programming interface (API) functions available. A JavaScript-compliant Web browser, such as Netscape Navigator, is necessary to run the JavaScript code.

Kermit: A file-transfer protocol developed at Columbia University and placed in the public domain that is used to transfer files between PCs and mainframe computers over standard telephone lines. Data is transmitted in variable-length blocks up to 96 characters in length, and each block is checked for transmission errors. Kermit detects transmission errors and initiates repeat transmissions automatically.

Large Internet Packet: Abbreviated LIP. A mechanism that allows the Novell NetWare internetwork packet size to be increased from the default 576 bytes, thus increasing throughput over bridges and

routers. LIP allows workstations to determine the packet size based on the largest packet supported by the router; the larger packet size is also supported by Ethernet and token ring networks.

Legacy system: A computer system, developed to solve a particular business need, which, due to the passage of time, has become obsolete. Legacy systems do not conform to the technical standards or performance standards of up-to-date systems. There is usually a requirement to maintain backward compatibility with or connections to legacy systems.

Link: On a Web page or a hypertext document, a connection between one element and another in the same or in a different document.

Local-area network: Abbreviated LAN. A group of computers and associated peripheral devices connected by a communications channel, capable of sharing files and other resources among several users.

Macro: A stored group of keystrokes or instructions that can automate a complex or repetitive sequence of application commands. Many of the major spreadsheet, word processing, and database programs let users create and edit macros to speed up operations. Some macros can incorporate control structures, such as DO/WHILE loops and IF/THEN branching statements.

Mail-aware application: Any application with the ability to send and receive e-mail. Applications in the document management, groupware, and workflow categories all use e-mail to interconnect users and help with the flow of information. This integration of e-mail is made possible in part by APIs such as Microsoft's Messaging API and Novell's Message Handling Service. Sometimes known as a message-enabled application.

Mailto: An HTML attribute that creates a link to an e-mail address. If a user clicks on the mail to link, the browser opens a window for composing an e-mail message to this address.

Mailer: A program used for sending and receiving e-mail.

Memory leak: A programming error that causes a program to request new areas of computer memory rather than reusing the memory already assigned to it. This causes the amount of memory in use by the program to increase as time goes on. In a worst case, the application may consume all available memory and stop the computer.

Mirror site: 1. A duplicate Web site. A mirror site contains the same information as the original Web site and reduces traffic on that site by providing a local or regional alternative. **2.** A duplicate data center. Large companies running mission-critical applications often mirror their entire data center so that the company can continue to function if the main center is hit by a natural disaster.

Network: A group of computers and associated peripheral devices connected by a communications channel capable of sharing files and other resources among several users. A network can range from a peer-to-peer network connecting a small number of users in an office or department, to a LAN connecting many users over permanently installed cables and dial-up lines, to a MAN or WAN connecting users on several networks spread over a wide geographic area.

Online analytical processing: Abbreviated OLAP. A category of software used to analyze historical business data to find previously hidden patterns. Analysts use OLAP software to view data in a multidimensional form, rather than in the more usual two-dimensional row and column format. In a multidimensional format, the intersection of important data is much more obvious, and data is easier to group and categorize.

Online service: A service that provides an online connection via modem for access to various services. Online services fall into these main groups:

- **Commercial services:** Services such as America Online charge a monthly membership fee for access to online forums, e-mail services, software libraries, and online conferences.

- **Internet:** The Internet is a worldwide network of computer systems and is not always easy to use, but the wealth of information available is staggering. The main problem for casual users is that there is no central listing of everything that is available.
- **Specialist databases:** Specific databases aimed at researchers can be accessed through online services such as Dow Jones News/Retrieval for business news and Lexis and Nexis for legal information and news archives.

Open architecture: A vendor-independent design that is publicly available and well understood within the industry.

Open Desktop: A graphical user interface from SCO that provides access to files and system utility functions on the desktop. Files, directories, and applications are represented by icons and displayed in windows.

Open source software: Any software package that includes the original source code from which the product was originally created. Open source software allows knowledgeable users to make changes to the way the software actually works, unlike products from mainstream software developers which never include the source code. And while this mainstream software is certainly configurable, it is basically a take-it-or leave- it package.

Packet: Any block of data sent over a network or communications link. Each packet may contain sender, receiver, and error-control information, in addition to the actual message, which may be data, connection management controls, or a request for a service. Packets may be fixed- or variable length, and they will be reassembled if necessary when they reach their destination. The actual format of a packet depends on the protocol that creates the packet; some protocols use special packets to control communications functions in addition to data packets.

Page: A single document available on the World Wide Web or on a corporate intranet. A page can contain any combination of text, graphics, animated graphics, audio, and video and can be static or dynamic.

Passive termination: A method used to terminate a Small Computer System Interface (SCSI) chain of devices. Passive termination is a simple termination method that works best with four or fewer devices on a SCSI daisy chain.

Password encryption: In certain operating systems, the password you enter to gain access to the system is not stored as ordinary text, but is encrypted, and this encrypted form is compared against the encrypted password stored on the server. If the two match, the logon continues; if not, the logon attempt is rejected.

Programming language: A language used to write a program that the computer can execute. Almost 200 programming languages exist. An example is the popular C language, which is well suited to a variety of computing tasks. With C, programmers can write anything from a device driver, to an application, to an operating system. Certain kinds of tasks, particularly those involving artificial intelligence (LISP or Prolog), process control (Forth), or highly mathematical applications (Fortran and APL), can benefit from a more specific language.

Proxy server: A software package running on a server positioned between an internal network and the Internet. The proxy server filters all outgoing connections so that they appear to be coming from the same machine, in an attempt to conceal the underlying internal network structure from any intruders. By disguising the real structure of the network, the proxy server makes it much more difficult for an intruder to mount a successful attack. A proxy server will also forward your requests to the Internet, intercept the response, and then forward the response to you at your network node. A system administrator can also regulate the external sites to which users can connect.

Query language: In a database management system, a programming language that allows a user to extract and display specific information from a database. Structured Query Language (SQL) is an international database query language that allows the user to issue high-level commands or statements, such as SELECT or INSERT, to create or modify data or the database structure.

Remote access: A workstation-to-network connection, made using a modem and a telephone line, that allows data to be sent and received over large distances. Remote access and authentication and security for such access is managed differently in different network operating systems.

Scripting: The process of invoking a script, written in a scripting language, from an HTML document on a Web site. Scripts can be written in a range of languages, including Perl, Tcl, REXX, JavaScript, JScript, or even Microsoft Visual Basic.

Search engine: A special Web site that lets you perform keyword searches to locate Web pages; To use a search engine, you enter one or more keywords or, in some cases, a more complex search string such as a Boolean expression. The search engine returns a list of matching Web pages, newsgroups, and FTP archives taken from its database, usually ranked in some way, that contain the expression you are looking for, along with a brief text description of the material. Searching this database is much faster than actually searching the Internet, but the accuracy and relevance of the information it contains depend on how often the data is updated and on the proportion of the Web that is actually searched for new content.

System administrator: The person charged with the responsibility of managing the system. In a very large system, the system administrator may in fact be several people or even a department; if you are running Linux on your system at home, you have to be your own system administrator.

Tag: An element in HTML used to annotate a document. A tag is text enclosed by angle brackets that tells the client Web browser how to display each part of the document. For example, the tag <H1> indicates the start of a level one heading, and the tag </H1> indicates the end of a level one heading.

Uniform Resource Locator: Abbreviated URL. An address for a resource on the Internet. URLs are used as a linking mechanism between Web pages and as a method for Web browsers to access Web pages. A URL specifies the protocol to be used to access the resource (such as HTTP or FTP), the name of the server where the resource is located (as in www.sybex.com), the path to that resource (as in /catalog), and the name of the document to open (/index.html).

Viewer: An application launched by a Web browser to view a file that the browser cannot display by itself. Sometimes called a helper application.

Web browser: A client application that lets you look at hypertext documents, follow links to other HTML documents, and download files on the Internet or on a corporate intranet. When you find something that interests you as you browse through a hypertext document, you can click on that object, and the system automatically takes care of accessing the Internet host that holds the document you requested; you don’t need to know the IP address, the name of the host system, or any other details. A Web browser will also display the graphics in a Web page, play audio and video clips, and execute small Java or ActiveX programs called applets, although certain older Web browsers may need helper, or plug-in, applications to perform some of these tasks.

Web page: Information placed on a Web server for viewing with a Web browser. A Web page can contain text, graphics, audio or video clips, and links to other Web pages.

Web server: A hardware and software package that provides services to client computers running Web browsers. Clients make requests in the form of HTTP messages; the server responds to these messages, returning Web pages or other requested documents to the client. Most Web servers run a version of Unix or Microsoft Windows NT Server.

World Wide Web: Abbreviated WWW, W3, or simply the Web. A huge collection of hypertext pages on the Internet. World Wide Web concepts were developed in Switzerland by the European Laboratory for Particle Physics (known as CERN), but the Web is not just a tool for scientists; it is one of the most flexible and exciting tools in existence. Hypertext links connect pieces of information (text, graphics, animation, audio, and video) in separate HTML pages located at the same or at different Internet sites, and you explore these pages and links using a Web browser such as Netscape Navigator or Microsoft Internet Explorer.