ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΟΣ
ΔΙΟΙΚΗΣΗΣ & ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΤΕΧΝΙΚΕΣ ΑΝΑΚΑΛΥΨΗΣ ΕΝΔΙΑΦΕΡΟΥΣΑΣ ΠΛΗΡΟΦΟΡΙΑΣ ΣΕ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

- **ΜΠΙΛΑΛΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ**
- **ΓΟΥΝΑΡΗΣ ΑΓΓΕΛΟΣ**
- **ΠΕΠΕΛΑΣΗΣ ΣΠΥΡΙΔΩΝ**

- **ΚΑΘΗΓΗΤΗΣ/ΕΠΟΠΤΗΣ: ΠΑΠΑΣΤΕΡΓΙΟΥ ΘΩΜΑΣ**

ΑΘΗΝΑ, 27/01/2016

ΠΡΟΛΟΓΟΣ

Η πληροφορία, θεωρείται στις μέρες μας αναγκαίος συντελεστής καθώς οι λειτουργίες της διοίκησης των επιχειρήσεων στηρίζονται όλο και περισσότερο στην αποτελεσματική χρησιμοποίηση της πληροφορίας, καθώς και στα συστήματα που την παρέχουν, δηλαδή τα Πληροφοριακά Συστήματα.

Με την έννοια Πληροφορίας προσδιορίζουμε ένα σύνολο σχετιζομένων και συνδεδεμένων συνιστωσών που μπορούν να λειτουργούν σαν μια ολότητα, προκειμένου να κατευθύνουν τα υποσυστήματα του ή άλλα εξαρτώμενα πληροφοριακά συστήματα στον επιθυμητό στόχο. Το πληροφοριακό σύστημα διαχειρίζεται πληροφορίες και έχει ως συστατικά στοιχεία άτομα, υλικό, λογισμικό και διαδικασίες που διέπονται από συγκεκριμένα πλαίσια.

Η σχέση μεταξύ του Πληροφοριακού Συστήματος και των τεχνολογιών πληροφορικής είναι εξαιρετικά σημαντική και αυτό γιατί οι τεχνολογίες πληροφορικής είναι εργαλεία και τεχνικές που υποστηρίζουν την ανάπτυξη των πληροφοριακών συστημάτων. Το λογισμικό, ο εξοπλισμός και οι τηλεπικοινωνίες είναι μερικά από τα εργαλεία και τεχνικές που χρησιμοποιεί το ΠΣ για να δημιουργήσει και να αποθηκεύσει πληροφορίες καθώς επίσης και για να δώσει τη δυνατότητα προσπέλασης σε αυτές.

Τα πληροφοριακά συστήματα χρησιμοποιούνται σε κάθε σύγχρονη επιχείρηση και διευκολύνουν τις λειτουργίες και αποφάσεις των οργανισμών αυτών. Διακρίνονται σε χειρογραφικά, όπου όλες οι λειτουργίες τους γίνονται δίχως τη χρήση Η/Υ και σε μηχανογραφημένα όπου ο Η/Υ είναι απαραίτητος για την ολοκλήρωση των συστημάτων και την εξόρυξη δεδομένων.

ΠΕΡΙΛΗΨΗ

Σκοπός της παρούσης μελέτης ήταν να παρουσιάσει και να περιγράψει τις τεχνικές ανακάλυψης μιας ενδιαφέρουσας πληροφορίας σε βάση δεδομένων. Η μεθοδολογία της εργασίας στηρίχθηκε στη συλλογή δευτερογενών δεδομένων, τα οποία συλλέχθηκαν μέσα από βιβλία άρθρα σε περιοδικά αλλά και μέσα από επίσημους διαδικτυακούς τόπους. Η εργασία ολοκληρώθηκε μέσα από τέσσερα κεφάλαια.

Το πρώτο κεφάλαιο μελέτησε τα πληροφοριακά συστήματα αποσαφηνίζοντας εννοιολογικά τον ορισμό της πληροφορίας και παρουσιάζοντας τα πληροφορικά συστήματα και συγκεκριμένα τα συστήματα επεξεργασίας δοσοληψιών, τα πληροφοριακά συστήματα διοίκησης, τα συστήματα υποστήριξης αποφάσεων, τα συστήματα υποστήριξης της εκτελεστικής εξουσίας και τα έμπειρα συστήματα.

Το δεύτερο κεφάλαιο εστίασε στην εισαγωγή στην εξόρυξη των δεδομένων και συγκεκριμένα στην εξέλιξη της εξόρυξης, στη διαδικασία της και τις τεχνικές της όπως είναι οι τεχνικές με βάση το δέντρο, και με βάση τους αλγόριθμους.

Το τρίτο κεφάλαιο αναφέρθηκε στο web wrappers και συγκεκριμένα στη δημιουργία και εκτέλεση των wrappers στα προβλήματα συντήρησής του και τέλος στη δημιουργία wrapper με βάση τη μάθηση.

Το τέταρτο και τελευταίο κεφάλαιο μελέτησε τις τεχνικές εξόρυξης μιας ενδιαφέρουσας πληροφορίας και συγκεκριμένα τον τρόπο εξόρυξης των δεδομένων, τις κύριες φάσεις που σχετίζονται με το σύστημα εξαγωγής των δεδομένων του ιστού και τέλος με τις συγκρίσεις των layer cake.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ

Σκοπός της παρούσης μελέτης είναι να παρουσιάσει και να περιγράψει τις τεχνικές ανακάλυψης μιας ενδιαφέρουσας πληροφορίας σε βάση δεδομένων. Η μεθοδολογία της εργασίας στηρίζεται στη συλλογή δευτερογενών δεδομένων, τα οποία συλλέχθηκαν μέσα από βιβλία άρθρα σε περιοδικά αλλά και μέσα από επίσημους διαδικτυακούς τόπους. Η εργασία ολοκληρώνεται μέσα από τέσσερα κεφάλαια. Το πρώτο κεφάλαιο μελετά τα πληροφοριακά συστήματα αποσαφηνίζοντας εννοιολογικά τον ορισμό της πληροφορίας και παρουσιάζοντας τα πληροφορικά συστήματα και συγκεκριμένα τα συστήματα επεξεργασίας δοσοληψιών, τα πληροφοριακά συστήματα διοίκησης, τα συστήματα υποστήριξης αποφάσεων, τα συστήματα υποστήριξης της εκτελεστικής εξουσίας και τα έμπειρα συστήματα.

Το δεύτερο κεφάλαιο εστιάζει στην εισαγωγή στην εξόρυξη των δεδομένων και συγκεκριμένα στην εξέλιξη της εξόρυξης, στη διαδικασία της και τις τεχνικές της όπως είναι οι τεχνικές με βάση το δέντρο, και με βάση τους αλγόριθμους.

Το τρίτο κεφάλαιο εστιάζει στο web wrappers και συγκεκριμένα στη δημιουργία και εκτέλεση των wrappers στα προβλήματα συντήρησής του και τέλος στη δημιουργία wrapper με βάση τη μάθηση

Το τέταρτο και τελευταίο κεφάλαιο μελετά τις τεχνικές εξόρυξης μιας ενδιαφέρουσας πληροφορίας και συγκεκριμένα τον τρόπο εξόρυξης των δεδομένων, τις κύριες φάσεις που σχετίζονται με το σύστημα εξαγωγής των δεδομένων του ιστού και τέλος με τις συγκρίσεις των layer cake.

Κεφάλαιο 1ο Πληροφοριακά συστήματα-Έννοια της πληροφορίας

1.1 Ορισμός της Πληροφορίας

Οι καλές πληροφορίες πρέπει να είναι κατάλληλες και να σχετίζονται με το πρόβλημα που εξετάζεται. Πρέπει επίσης να είναι έγκυρες. Για παράδειγμα, οι πληροφορίες από την έρευνα για την μπίρα Buckler (χωρίς οινόπνευμα) της Αθηναϊκής Ζυθοποιίας Α.Ε. θα ήταν άχρηστες αν δίνονταν δύο χρόνια μετά την απόσυρση του προϊόντος (Durbin, 1997).

Οι καλές πληροφορίες πρέπει, επίσης, να είναι ακριβείς και τελικά οι καλές πληροφορίες μειώνουν την αβεβαιότητα, η οποία δημιουργείται από την έλλειψη πληροφοριών για μια συγκεκριμένη περιοχή ενδιαφέροντος. Στο παράδειγμα της Αθηναϊκής Ζυθοποιίας Α.Ε, για να εκπληρώνει αυτά τα κριτήρια η έρευνα πληροφοριών, θα πρέπει να βοηθά το διευθυντή του μάρκετινγκ να απαντήσει στο ερώτημα: "Γιατί οι άνθρωποι δεν αγοράζουν την Buckler με τον τρόπο που νομίζαμε ότι θα το έκαναν;".

Εντούτοις, ακόμη και οι καλές πληροφορίες είναι σχετικά άχρηστες, χωρίς τις γνώσεις που προέρχονται από την ανάλυση και την ερμηνεία τους. Σήμερα, τα στελέχη των επιχειρήσεων κατακλύζονται, αν μη τι

άλλο, από πληροφορίες για τις πρακτικές των ανταγωνιστών, για τις αγοραστικές συνήθειες των καταναλωτών, για τη λεπτομερειακή ανάλυση των μηχανών και για πολλά άλλα σχετικά θέματα. Έτσι, ο ρόλος της τεχνολογίας πληροφοριών οργάνωσης δεν είναι μόνο να συλλέγει και να μεταβιβάζει περισσότερες (ή ακόμη καλύτερης ποιότητας) πληροφορίες, αλλά να εφοδιάσει τα στελέχη με τις απαραίτητες γνώσεις, μέσα από την ανάλυση και την ερμηνεία για το τι ακριβώς συμβαίνει στην επιχείρησή τους.

1.2 Πληροφοριακά συστήματα

Η τεχνολογία των πληροφοριών αναφέρεται στις διαδικασίες, τις πρακτικές ή τα συστήματα που διευκολύνουν την επεξεργασία και τη μεταφορά πληροφοριών (Kennedy, 1997). Αναμφίβολα, σήμερα οι περισσότεροι είναι πολύ εξοικειωμένοι με τα σύγχρονα συστατικά της τεχνολογίας των πληροφοριών. Για παράδειγμα, μπορεί να χρησιμοποιούν προσωπικό ηλεκτρονικό υπολογιστή και από τη δουλειά την οποία κάνουν. να είναι εξοικειωμένοι με τα πληροφοριακά συστήματα διοίκησης. Πιθανόν χρησιμοποιούν κυψελοειδή τηλέφωνα, τηλεομοιοτυπία και τα όλο και πιο διαδεδομένα συστήματα ηλεκτρονικού ταχυδρομείου και ταχυδρομείου φωνής. Αυτού του είδους οι τεχνολογίες των πληροφοριών άλλαξαν δραματικά τον τρόπο με τον οποίο οι άνθρωποι κάνουν τις δουλειές τους και τον τρόπο με τον οποίο διοικούνται οι επιχειρήσεις.

Ένα πληροφοριακό σύστημα μπορεί να οριστεί ως ένα σύνολο ανθρώπων, δεδομένων, τεχνολογίας και οργανωτικών μεθόδων που δουλεύουν μαζί για να συλλέξουν. να επεξεργαστούν. να αποθηκεύουν και να μεταβιβάσουν πληροφορίες για να στηρίξουν τη λήψη αποφάσεων και τον έλεγχο: Ειδικά, θα εστιάσουμε την ανάλυση στα πληροφοριακά συστήματα διοίκησης, τα οποία είναι συστήματα που στηρίζουν τη λήψη αποφάσεων και τον έλεγχο από τη διεύθυνση των επιχειρήσεων.

Τα πληροφοριακά συστήματα δεν είναι απλώς οι ηλεκτρονικοί υπολογιστές. Συνήθως, το πληροφοριακό σύστημα περιλαμβάνει και την επιχείρηση ή σημαντικά μέρη της, όπως τους εργαζομένους που εισάγουν δεδομένα στο σύστημα και παίρνουν πίσω την εκροή του. Τα στελέχη επιχειρήσεων είναι (ή θα έπρεπε να είναι) μέρος του πληροφοριακού συστήματος, αφού το πληροφοριακό σύστημα είναι σχεδιασμένο για να υπηρετεί τις ειδικές ανάγκες τους για πληροφορίες.

Τα πληροφοριακά συστήματα διακρίνονται στα εξής βασικά είδη :

Α) Συστήματα Επεξεργασίας Δοσοληψιών (Transaction Processing Systems – T.P.S): Μια δοσοληψία είναι ένα συμβάν που επηρεάζει την επιχείρηση.

Τα συστήματα επεξεργασίας δοσοληψιών αυτοματοποιήθηκαν οι διαδικασίες εκείνες που επαναλαμβάνονται. Ως παραδείγματα μπορεί να αναφερθούν η χρήση των Η/Υ για τους παρακρατούμενους φόρους (Φ.Π.Α., Ο.Γ.Α., κ.ά.), για την επεξεργασία επιταγών πληρωτέων λογαριασμών, κ.ά. Τα συστήματα επεξεργασίας δοσοληψιών μπορεί να

έχουν πέντε χρήσεις. Έτσι αυτά χρησιμοποιούνται:

1. Για την ταξινόμηση δεδομένων που βασίζονται στα κοινά χαρακτηριστικά μιας ομάδας (όπως, π.χ., να βρουν τους εργαζομένους στο τμήμα πωλήσεων, με πενταετή υπηρεσία).
2. Για υπολογισμούς ρουτίνας (όπως το να περνούν στον Η/Υ τις καθαρές αμοιβές μετά από τους φόρους και τις κρατήσεις για κάθε εργαζόμενο).
3. Για την ταξινόμηση σε ομάδες (για παράδειγμα, συγκέντρωση τιμολογίων κατά ομάδες ανάλογα με τον ταχυδρομικό τομέα, ώστε να γίνεται πιο αποδοτικά η διανομή τους).
4. Για συνοπτικούς λογαριασμούς (για παράδειγμα, συνοπτικό λογαριασμό για κάθε προϊστάμενο τμήματος, που δείχνει τις μέσες μισθολογικές δαπάνες του τμήματός του σε σύγκριση με τα άλλα τμήματα).
5. Τέλος, τα συστήματα επεξεργασίας δοσοληψιών μπορεί να χρησιμοποιηθούν για αποθήκευση (για παράδειγμα, αποθήκευση πληροφοριών για τις μισθολογικές καταστάσεις τα τελευταία πέντε χρόνια).

Β. Πληροφοριακά Συστήματα Διοίκησης (Management Information Systems - M.I.5.): Ένα πληροφοριακό σύστημα διοίκησης στηρίζει τη λήψη αποφάσεων των στελεχών των επιχειρήσεων, παράγοντας πρότυπες, συνοπτικές εκθέσεις σε τακτική βάση. Τα συστήματα αυτά παράγουν εκθέσεις για μακροπρόθεσμους στόχους, σε σύγκριση με τα συστήματα επεξεργασίας δοσοληψιών που ασχολούνται με διαδικασίες ρουτίνας.

Γ. Συστήματα Υποστήριξης Αποφάσεων (Decision Support systems - 0.5.5.): Τα συστήματα υποστήριξης αποφάσεων βοηθούν τα στελέχη των επιχειρήσεων στη λήψη των αποφάσεων. Τα συστήματα αυτά συνδυάζουν δεδομένα, επεξεργασμένα αναλυτικά πρότυπα και ένα φιλικό για το χρήστη λογισμικό σε ένα ενιαίο ισχυρό σύστημα, που μπορεί να υποστηρίξει ημιδομημένα ή μη δομημένα προβλήματα. Με άλλα λόγια, αυτά τα συστήματα μπορεί να βοηθήσουν τα στελέχη επιχειρήσεων να πάρουν αποφάσεις για μη δομημένα προβλήματα.

Ένα σύστημα υποστήριξης αποφάσεων (O.S.S.) διαφέρει από ένα πληροφοριακό σύστημα διοίκησης (M.I.S.) σε πολλά σημεία.

Έτσι, τα συστήματα υποστήριξης αποφάσεων ασχολούνται με προβλήματα που δεν είναι προγραμματισμένα, τα οποία όμως χρειάζονται την κριτική παρέμβαση του στελέχους, ενώ τα πληροφοριακά συστήματα διοίκησης ασχολούνται βασικά με προβλήματα που είναι προγραμματισμένα και με αποφάσεις ρουτίνας. Επιπλέον, ένα σύστημα υποστήριξης αποφάσεων δεν στηρίζεται μόνο στις εσωτερικές πληροφορίες από το σύστημα επεξεργασίας δοσοληψιών, όπως στηρίζεται τυπικά το πληροφοριακό σύστημα διοίκησης. Αντίθετα, ένα σύστημα υποστήριξης αποφάσεων είναι έτσι δομημένο ώστε να απορροφά στην ανάλυση νέες εξωτερικές πληροφορίες.

Δ. Συστήματα Υποστήριξης της Εκτελεστικής Εξουσίας (Executive Support systems - E.S.S.): Τα συστήματα υποστήριξης της εκτελεστικής εξουσίας είναι πληροφοριακά συστήματα σχεδιασμένα για να βοηθούν την εκτελεστική εξουσία ανώτερου επιπέδου να αποκτά, να χειρίζεται και να χρησιμοποιεί τις πληροφορίες που χρειάζεται, προκειμένου να διατηρεί τη συνολική αποτελεσματικότητα της επιχείρησης. Αυτά τα συστήματα εστιάζονται συχνά στο να παρέχουν στην ανώτερη διεύθυνση πληροφορίες για τη λήψη στρατηγικών αποφάσεων.

Οι εκτελεστικοί μάνατζερ χρησιμοποιούν, επίσης, τα συστήματα υποστήριξης της εκτελεστικής εξουσίας για να ανιχνεύσουν το περιβάλλον της επιχείρησης. Για παράδειγμα, πολλές πληροφορίες είναι διαθέσιμες σε ηλεκτρονικές τράπεζες δεδομένων, στις οποίες περιλαμβάνονται πληροφορίες για πολλές επιχειρήσεις της χώρας μας. Τέλος, ένα σύστημα υποστήριξης της εκτελεστικής εξουσίας επιτρέπει στους εκτελεστικούς μάνατζερ να έχουν άμεση πρόσβαση στα δεδομένα. Χρησιμοποιώντας τα τερματικά τους και τις τηλεφωνικές γραμμές τους, οι εκτελεστικοί μάνατζερ μπορούν να χρησιμοποιήσουν ένα σύστημα υποστήριξης της εκτελεστικής εξουσίας για να μπαίνουν άμεσα στα αρχεία δεδομένων της εταιρείας, ώστε να παίρνουν ειδικές πληροφορίες για τις οποίες μπορεί να ενδιαφέρονται, χωρίς να περιμένουν να τους τις συγκεντρώσουν άλλοι.

Ε. Έμπειρα Συστήματα (Expert Systems - E.S): Ένα έμπειρο σύστημα είναι ένα πληροφοριακό σύστημα, στο οποίο τα προγράμματα ηλεκτρονικού υπολογιστή αποθηκεύουν γεγονότα και κανόνες (αποκαλούνται συχνά βάση γνώσεων), ώστε να αντιγράφουν τις ικανότητες και τις αποφάσεις ανθρώπων που είναι έμπειροι. Για παράδειγμα, μια πρώιμη εφαρμογή εντόπιζε τα κριτήρια ενός συμβούλου επενδύσεων με βάση τα οποία σύστηνε επενδύσεις σε πελάτες που ήταν σε διάφορες δημογραφικές κατηγορίες και σε ποικίλες κατηγορίες ως προς την τάση ανάληψης κινδύνων. Κατόπιν αυτές οι παρατηρήσεις χρησιμοποιούνταν για να αναπτυχθεί ένα πρόγραμμα ηλεκτρονικού υπολογιστή, το οποίο αναπαρήγαγε τις περισσότερες από τις αποφάσεις επενδύσεων τις οποίες θα είχε κάνει ο (έμπειρος) σύμβουλος επενδύσεων. Τα έμπειρα συστήματα χρησιμοποιούνται σε όλους τους τομείς επιχειρήσεων, από την παραγωγή μέχρι το μάρκετινγκ και το χρηματοοικονομικό τομέα . Ωστόσο όλο και περισσότερο ,μια από τις πιο προσβεβλημένες χρήσεις, είναι στο χρηματοοικονομικό τομέα και στις επενδύσεις.

1.3 Πληροφοριακά συστήματα και ασφάλεια

Η σημασία και ο ρόλος της τεχνολογίας στην οικονομία και στην κοινωνία έχει γίνει εξαιρετικά σημαντικός την τελευταία δεκαετία. Οι άνθρωποι, οι επιχειρήσεις και οι κυβερνήσεις έχουν αγκαλιάσει με ενθουσιασμό την τεχνολογία καθώς κατανοούν ότι η τελευταία έχει πολλά να προσφέρει. Με το όφελος από μια ιστορική προοπτική της σταδιακής εξέλιξης των τεχνολογιών οι άνθρωποι συχνά ενσωματώνουν τις διάφορες

τεχνολογίες στην κοινωνία και παράλληλα τις διαμορφώνουν μέσα στη ζωή τους για το λόγο ότι ο ρόλος της πληροφορίας και της επικοινωνίας τεχνολογιών στη διαμόρφωση της ανθρώπινης δραστηριότητας στην ιδιωτική και επαγγελματική μας ζωή έχει γίνει όλο και πιο έντονος.

Σήμερα τα πληροφοριακά συστήματα αποτελούν το θεμέλιο για την ομαλή λειτουργία των επιχειρήσεων. Σε πολλές βιομηχανίες, η επιβίωση και ακόμη και η ύπαρξη είναι δύσκολη χωρίς την εκτεταμένη χρήση αυτής της τεχνολογίας. Τα πληροφοριακά συστήματα έχουν καταστεί ουσιώδης βοήθεια για τους οργανισμούς που λειτουργούν σε μια παγκόσμια οικονομία. Οι επιχειρήσεις προσπαθούν να γίνουν πιο ανταγωνιστικές και αποτελεσματικές μέσα από το μετασχηματισμό τους σε ψηφιακές επιχειρήσεις όπου σχεδόν όλες οι βασικές επιχειρηματικές διεργασίες και οι σχέσεις με τους πελάτες, τους προμηθευτές και τους εργαζόμενους έχουν ενεργοποιηθεί ψηφιακά.

Οι επιχειρήσεις χρησιμοποιούν σήμερα τα συστήματα πληροφοριών για την επίτευξη έξι κύριων στόχων: 1) την επιχειρησιακή αριστεία 2) τα νέα προϊόντα, 3) τις υπηρεσίες και τα επιχειρηματικά μοντέλα, 4)την οικειότητα με τον πελάτη /προμηθευτή, 5) τη βελτιωμένη λήψη αποφάσεων και 6) το ανταγωνιστικό πλεονέκτημα (Galliers, and Leidner, 2014).

1.4 Βασικές έννοιες θεωρίας συστημάτων

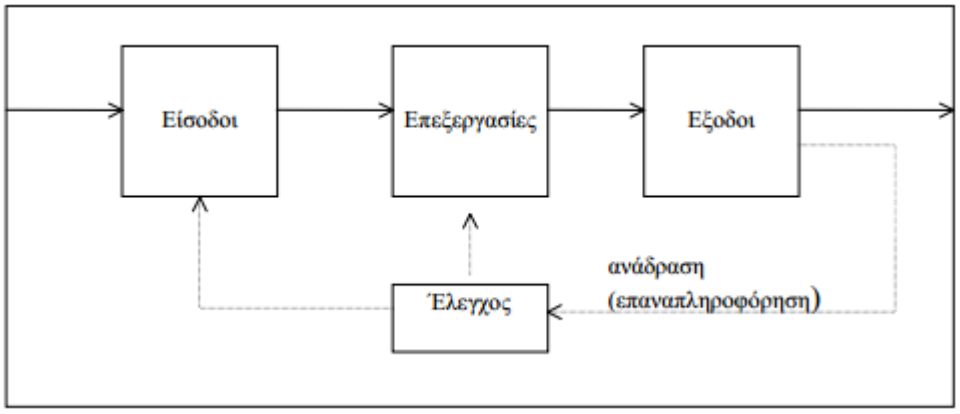
Τα συστήματα πληροφοριών είναι ένα διεπιστημονικό πεδίο και πολλές διαφορετικές θεωρίες και προοπτικές έχουν χρησιμοποιηθεί για να τα περιγράψουν. Κλάδοι όπως: η επιστήμη των υπολογιστών, η επιχειρησιακή έρευνα, η κοινωνιολογία, η οικονομία, η ψυχολογία και η επιστήμη της διαχείρισης, όλα αυτά μαζί συμβάλλουν στην κατανόησή μας για το πώς μπορούν να εφαρμοστούν και να χρησιμοποιηθούν στα πληροφοριακά συστήματα. Σε γενικές γραμμές, το πεδίο των πληροφοριακών συστημάτων μπορεί να υποδιαιρεθεί σε τεχνικές και συμπεριφορικές προσεγγίσεις. Θεωρητικά ένα σύστημα είναι το θεμέλιο για μια πραγματικά γενική κατανόηση του κόσμου.

Το «σύστημα» ως έννοια είναι καθολική και αναφέρεται σε κάθε «πράγμα» που υπάρχει στη φύση ανεξάρτητα από τις ιδιαίτερες ιδιότητες, την καταγωγή ή την ουσία του. Έτσι παρέχει μια ενοποιητική ή σύνθετη δύναμη για την αντιμετώπιση της διασπάσεως ή των αναλυτικών πτυχών της επιστήμης και της εμπειρικής γνώσης. Με τη χρήση αυτής της αναλογίας μπορεί κανείς να εξετάσει πτυχές των πραγμάτων και των φαινομένων που ισχύουν για όλα τα πράγματα και τα φαινόμενα. Μπορεί να χρησιμοποιηθεί για να μεταφράσει και να ενοποιήσει τις έννοιες σε πολλά κατακερματισμένα και συχνά αποξενωμένα πεδία γνώσης (Pinedo, 2012).

Η προσέγγιση αυτή προσδιόρισε και καθιέρωσε ένα νέο τρόπο σκέψης που ονομάστηκε θεωρία συστημάτων. Η εν λόγω προσέγγιση έχει άμεση σύνδεση με την αντίληψη που έχουμε για τον κόσμο. Νέοι κλάδοι των επιστημών δημιουργήθηκαν στηριζόμενοι στη θεωρία συστημάτων: οι Επιχειρησιακές Έρευνες, η Διοίκηση Επιχειρήσεων και η Ανάλυση Συστημάτων είναι ορισμένοι από αυτούς. Το κεφάλαιο αυτό αναπτύσσει ορισμένες έννοιες της θεωρίας των συστημάτων. Οι έννοιες αυτές είναι οι

ακόλουθες (Alter, 2013):

1. Σύστημα (system) Ένα σύστημα είναι ένα ειδικό μοντέλο σύμφωνα με το οποίο όλα τα πράγματα που περιέχονται σε αυτό (όλες οι συνιστώσες του συστήματος) είναι μεταβατικά συνεκτικές. Δηλαδή όλα αυτά άμεσα ή έμμεσα συνδέονται μεταξύ τους και σχηματίζουν ένα συνεκτικό σύνολο. Κάθε σύστημα περιλαμβάνει εισόδους, εξόδους και επεξεργασίες και διέπεται από ένα περιβάλλον από το οποίο διαχωρίζεται από ένα όριο. Τις περισσότερες φορές υπάρχει κάποιος ο οποίος είναι υπεύθυνος για την λήψη αποφάσεων σχετικών με το σύστημα (Βλέπε Σχήμα 1.1).



Σχήμα 1.1 Τα στοιχεία που απαρτίζουν ένα σύστημα

2. Είσοδος (Input), είναι τα δεδομένα που συλλέγονται από το σύστημα και τα αποτελέσματα είναι τα δεδομένα που κατευθύνονται σε αυτό. Ο όρος μπορεί επίσης να χρησιμοποιηθεί ως μέρος μιας δράσης για να "εκτελέσει μια λειτουργία εισόδου ή εξόδου. Η έξοδος από αυτές τις συσκευές είναι η είσοδος για τον υπολογιστή. Με την ίδια διαδικασία οι συσκευές που συνδέονται σε ένα pc λαμβάνουν ως σήματα εισόδου τα δεδομένα που ο υπολογιστής εξάγει. Κατόπιν, μετασχηματίζουν τα εν λόγω σήματα σε απεικονίσεις τις οποίες οι χρήστες δύναται να δουν και να διαβάσουν.
3. Η ανάδραση ή ανατροφοδότηση (feedback) είναι οι πληροφορίες από το σύστημα που χρησιμοποιείται για να κάνει οι αλλαγές στις δραστηριότητες είσοδο ή τη μεταποίηση. Για παράδειγμα, λάθη ή προβλήματα θα μπορούσαν να καταστήσουν αναγκαία τη διόρθωση δεδομένων εισόδου ή την αλλαγή μιας διαδικασίας.

1.5 Πληροφοριακά συστήματα

Τα δεδομένα είναι τα ρεύματα των πρώτων στοιχείων που αντιπροσωπεύουν τα γεγονότα που συμβαίνουν σε οργανισμούς ή το φυσικό περιβάλλον πριν οργανωθούν και τοποθετηθούν σε μια μορφή όπου οι άνθρωποι μπορούν να τα κατανοήσουν και να τα χρησιμοποιήσουν.

Τα δεδομένα προκειμένου να είναι χρήσιμα χρειάζεται να διαθέτουν τα παρακάτω χαρακτηριστικά τα οποία προσδιορίζουν την ποιότητά τους (Sousa, and Oz, 2014):

1. **Ακριβή**: Οι πληροφορίες πρέπει να είναι αρκετά ακριβείς για τη χρήση για την οποία πρόκειται να τεθούν. Η ακρίβεια είναι σημαντική. Για παράδειγμα, εάν οι στατιστικές της κυβέρνησή με βάση την τελευταία απογραφή δείχνουν μεγάλη αύξηση των γεννήσεων εντός της ίδιας περιοχής, τα σχέδια μπορούν να γίνουν για την κατασκευή σχολείων με τις κατασκευαστικές εταιρείες να μπορούν να επενδύσουν σε νέα συγκροτήματα κατοικιών. Σε αυτές τις περιπτώσεις δεν μπορεί να αποσβεστεί οποιαδήποτε επένδυση.
2. **Πλήρη**: Οι πληροφορίες θα πρέπει να περιέχουν όλα τα στοιχεία που απαιτούνται από το χρήστη. Διαφορετικά, μπορεί να μην είναι χρήσιμες ως βάση για τη λήψη απόφασης. Για παράδειγμα, αν ένας οργανισμός τροφοδοτείται με πληροφορίες σχετικά με τις δαπάνες της προμήθειας στόλου των αυτοκινήτων για τη δύναμη των πωλήσεων και το κόστος επισκευής και συντήρησης δεν περιλαμβάνονται, κατόπιν κοστολόγησης με βάση τις πληροφορίες που παρέχονται θα πρέπει να υποτιμηθεί σημαντικά.
3. **Σχετικά**: οι σχετικές πληροφορίες είναι σημαντικές για τη λήψη αποφάσεων
4. **Έγκαιρα**: να είναι διαθέσιμα όταν τα χρειάζεται η οργάνωση.

Ένας κοινός ορισμός της πληροφορίας είναι ότι αποτελεί τα "δεδομένα" που έχουν υποστεί επεξεργασία ώστε να είναι σημαντικά. Αυτό απαιτεί μια διαδικασία που χρησιμοποιείται για την παραγωγή των πληροφοριών η οποία περιλαμβάνει τη συλλογή δεδομένων και στη συνέχεια την υποβολή τους σε μία διαδικασία μετασχηματισμού με σκοπό τη δημιουργία των πληροφοριών. Μερικοί παραδείγματα των πληροφοριών περιλαμβάνουν: την πρόβλεψη των πωλήσεων ή των οικονομικών καταστάσεων.

Από τεχνική άποψη ένα πληροφοριακό σύστημα συλλέγει, αποθηκεύει και διαχέει πληροφορίες από το περιβάλλον ενός οργανισμού και τις εσωτερικές ενέργειες για τη στήριξη οργανωτικών λειτουργών και τη λήψη αποφάσεων, την επικοινωνία, το συντονισμό, τον έλεγχο, την ανάλυση και οπτικοποίηση. Τα πληροφοριακά συστήματα μετατρέπουν τα ανεπεξέργαστα δεδομένα σε χρήσιμες πληροφορίες μέσα από τρεις βασικές δραστηριότητες: την εισαγωγή, την επεξεργασία και την παραγωγή. Από μια επιχειρηματική προοπτική, ένα πληροφοριακό σύστημα παρέχει μια λύση σε ένα πρόβλημα ή πρόκληση που αντιμετωπίζει μια επιχείρηση και παρέχει πραγματική οικονομική αξία για την εταιρεία.

Ένα Π.Σ. δύναται να είναι είτε χειρωνακτικό είτε να στηρίζεται στην χρήση ηλεκτρονικού υπολογιστή. Επίσης, ένα πληροφοριακό σύστημα εξαρτάται από την λειτουργία ηλεκτρονικού υπολογιστή και αξιοποιεί την τεχνολογία του προκειμένου να υλοποιήσει μια σειρά από στόχους. Επιπλέον, ένα Π.Σ. δύναται να είναι τυπικό ή άτυπο. Τα τυπικά συστήματα στηρίζουν την ύπαρξή τους στις διαδικασίες (που καθιερώνονται και που γίνονται αποδεκτές από την οργανωτική πρακτική) για τη συλλογή, την αποθήκευση, το χειρισμό και την πρόσβαση των στοιχείων προκειμένου να ληφθούν οι πληροφορίες. Τα τυπικά συστήματα δεν είναι απαραίτητο να μηχανογραφηθούν, αλλά σήμερα είναι συνήθως. Τα άτυπα συστήματα πληροφοριών υπάρχουν επίσης μέσα σε μια οργάνωση (διαπροσωπική δικτύωση, τα νέα στο γραφείο και το

κουτσομπολιό ή η ανταλλαγή μηνυμάτων μεταξύ φίλων με το ηλεκτρονικό ταχυδρομείο. Παρακάτω αναλύονται οι δραστηριότητες ενός Π.Σ. Τα δεδομένα συλλέγονται από διάφορες πηγές (Thimm, and Rasmussen, 2013):

1. Από εσωτερικές πηγές (internal sources) - π.χ. δεδομένα σχετικά με τις παραγγελίες που είναι έτοιμες προς αποστολή.
2. Από εξωτερικές πηγές (external sources) - π.χ. δεδομένα σχετικά με τις παραγγελίες των πελατών
3. Από το περιβάλλον - π.χ. δεδομένα που συλλέγονται από εταιρίες δημοσκοπήσεων.

Τα δεδομένα είναι μόνο τα ακατέργαστα γεγονότα, το υλικό για τη λήψη των πληροφοριών. Τα συστήματα πληροφοριών χρησιμοποιούν τα στοιχεία (δεδομένα) που αποθηκεύονται στις βάσεις δεδομένων υπολογιστών για να παρέχουν τις αναγκαίες πληροφορίες. Μια βάση δεδομένων είναι μια οργανωμένη συλλογή των αλληλένδετων στοιχείων που απεικονίζουν μια σημαντική πτυχή των δραστηριοτήτων μιας εταιρείας.

Τα δεδομένα αποτελούνται από πρωτογενή γεγονότα, όπως τα ονόματα των πελατών και οι διευθύνσεις. Οι πληροφορίες είναι μια συλλογή των γεγονότων που οργανώνονται με τέτοιο τρόπο ώστε να έχει μεγαλύτερη αξία πέρα από τα ίδια τα γεγονότα. Για παράδειγμα, μια βάση δεδομένων με τα ονόματα των πελατών και των αγορών θα μπορούσε να παράσχει πληροφορίες σχετικά με τα δημογραφικά στοιχεία της αγοράς μιας εταιρείας, οι τάσεις των πωλήσεων, και την αφοσίωση των πελατών / κύκλου εργασιών (Thimm, and Rasmussen, 2013).

Οι πόροι που περιλαμβάνει ένα πληροφοριακό σύστημα είναι οι εξής: α) ανθρώπινοι πόροι (τελικοί χρήστες, ειδικοί της πληροφορικής) β) υλικοί πόροι (συσκευές οι οποίες έχουν ως σκοπό την διαδικασία εισαγωγής, επεξεργασίας και αποθήκευσης των δεδομένων) γ) πόροι λογισμικού (προγράμματα και διαδικασίες) και δ) πόροι δεδομένων (βάσεις δεδομένων κλπ).

Η αναδυόμενη υπερ-ανταγωνιστική εποχή κατά τις τελευταίες δεκαετίες έχει αυξήσει την ανάγκη της για τη χρήση συστημάτων πληροφόρησης και τεχνολογίας στη διαχείριση των ανθρώπινων πόρων για την ανταγωνιστικότητα. Η επανάσταση στην τεχνολογία των πληροφοριών είναι πλήρης και γρήγορα επαναπροσδιορίζει τον τρόπο που τα πράγματα που γίνονται σχεδόν σε κάθε τομέα της ανθρώπινης δραστηριότητας.

Ανθρώπινοι πόροι και η τεχνολογία των πληροφοριών αποτελούν δύο στοιχεία όπου πολλές επιχειρήσεις αναζητούν για να τα χρησιμοποιήσουν ως στρατηγικά όπλα για να έχουν ανταγωνιστικό πλεονέκτημα. Επιπλέον, η συμμετοχή των ανθρώπων είναι έντονη είτε είναι τελικοί χρήστες είτε ειδικοί της πληροφορικής δύναται να περιλαμβάνει μηχανικούς, υπάλληλοι, λογιστές, διοικητικοί, κλπ. Οι ειδικοί της πληροφορικής δημιουργούν και χρησιμοποιούν τα Π.Σ. και οι οποίοι περιλαμβάνουν τους αναλυτές συστημάτων, τους προγραμματιστές, τους χειριστές ηλεκτρονικών υπολογιστών, κλπ (Yan, and Ma, 2014).

Στους υλικούς πόρους ανήκουν: α) το υλικό δηλαδή τα συστήματα ηλεκτρονικών υπολογιστών τα οποία αποτελούνται από κεντρική μονάδα επεξεργασίας, τα περιφερειακά (πληκτρολόγιο, οθόνη, εκτυπωτής, κ.λ.π) και τα δίκτυα τηλεπικοινωνιών, β) τα μέσα που χρησιμοποιούνται για την

αποθήκευση δεδομένων (χαρτί, μαγνητικές ταινίες, σκληροί δίσκοι, κλπ).

Οι Πόροι λογισμικού αποτελούν ένα κομμάτι της επιχείρησης το οποίο είναι σπουδαίας σημασίας. Αυτό το τμήμα περιλαμβάνει τα εξής: α) το λογισμικό συστήματος που τσεκάρει και ενθαρρύνει τις λειτουργίες του ηλεκτρονικού υπολογιστή π.χ τα λειτουργικά συστήματα, β) το λογισμικό εφαρμογών το οποίο προσφέρει στον τελικό χρήστη μια ευελιξία να επεξεργαστεί και να χειριστεί ένα ορισμένο πρόβλημα (λχ προγράμματα ανάλυσης πωλήσεων, προγράμματα μισθοδοσίας, επεξεργαστές κειμένου), γ) τις διαδικασίες δηλαδή υποδείξεις προς τους ανθρώπους που χρησιμοποιούν το Π.Σ. λ.χ. οδηγίες συμπλήρωσης μίας φόρμας, ενός προγράμματος (Yan, and Ma, 2014).

Τα περισσότερα από τα δεδομένα που συλλαμβάνονται από τα πληροφοριακά συστήματα σχετίζονται με τις δραστηριότητες του ίδιου του οργανισμού και χρησιμεύουν για την παραγωγή των εσωτερικών πληροφοριών. Αλλά σε μια ολοένα και πιο ανταγωνιστική αγορά, μια επιχείρηση πρέπει να έχει πρόσβαση σε περισσότερες εξωτερικές πληροφορίες. Ως εκ τούτου, είναι σημαντικό να σημειωθεί ότι οι ιθύνοντες χρειάζονται τόσο την εσωτερική πληροφόρηση σχετικά με την οργάνωσή τους όσο και τις εξωτερικές πληροφορίες σχετικά με το επιχειρησιακό τους περιβάλλον.

Τα δεδομένα δύναται να έχουν διάφορες μορφές (κείμενο, εικόνα, ήχος) και οργανώνονται σε (Yan, and Ma, 2014):

1. Βάσεις δεδομένων που αποθηκεύουν και διαχειρίζονται οργανωμένα δεδομένα,
2. Βάσεις προτύπων που αποθηκεύουν μαθηματικά και λογικά πρότυπα τα οποία περιέχουν σχέσεις, υπολογισμούς και αναλυτικές τεχνικές και τέλος
3. Βάσεις γνώσεων που αποθηκεύουν γεγονότα και κανόνες για διάφορα προβλήματα.

Ένα πληροφοριακό σύστημα αποτελεί ένα συνδυασμό της διαχείρισης, της οργάνωσης και των τεχνολογικών στοιχείων. Η διάσταση της διαχείρισης των πληροφοριακών συστημάτων περιλαμβάνει την ηγεσία, τη στρατηγική και τη συμπεριφορά της διαχείρισης. Οι διαστάσεις της τεχνολογίας επίσης αποτελούνται από το υλικό του υπολογιστή, το λογισμικό, την τεχνολογία διαχείρισης δεδομένων και την τεχνολογία δικτύωσης/ τηλεπικοινωνιών (συμπεριλαμβανομένου του Διαδικτύου). Η οργάνωση και διάσταση των συστημάτων πληροφοριών συνεπάγεται την ιεραρχία της οργάνωσης, λειτουργίας των ειδικοτήτων και των επιχειρηματικών διαδικασιών.

Τα πληροφοριακά συστήματα αποτελούν τον παράγοντα εκείνο που συντελεί στην καλή και δημιουργική συνεργασία ανθρώπινου δυναμικού, στοιχείων, τεχνολογιών πληροφορίας, διεργασιών και επικοινωνιών. Σήμερα τα πληροφοριακά συστήματα ως μάθημα (κατεύθυνση) διδάσκονται είτε σε προπτυχιακό είτε σε μεταπτυχιακό στάδιο.

Έξι λόγοι για τους οποίους τα συστήματα πληροφόρησης είναι τόσο σημαντική για τις επιχειρήσεις σήμερα περιλαμβάνουν τα εξής (Thimm, and Rasmussen, 2013):

- Τη λειτουργική αρτιότητα
- Τα νέα προϊόντα, υπηρεσίες και επιχειρηματικά μοντέλα
- Την οικειότητα τόσο με τον πελάτη όσο και με τον προμηθευτή

- Τη βελτιωμένη λήψη αποφάσεων
- Το ανταγωνιστικό πλεονέκτημα
- Την επιβίωση

Παρά το γεγονός ότι κάθε μία από τις κύριες λειτουργίες των επιχειρήσεων έχει το δικό της σύνολο των επιχειρηματικών διαδικασιών, πολλές άλλες επιχειρηματικές διαδικασίες είναι διασυνοριακά λειτουργικές, όπως η εκτέλεση παραγγελιών. Τα συστήματα πληροφοριών μπορούν να βοηθήσουν τους οργανισμούς να επιτύχουν μεγαλύτερη αποδοτικότητα, αυτοματοποιώντας μέρη αυτών των διαδικασιών, είτε βοηθώντας τον επανασχεδιασμό των επιχειρήσεων και τον εξορθολογισμό τους. Οι επιχειρήσεις μπορούν να γίνουν πιο ευέλικτες και αποδοτικές συντονίζοντας στενά τις επιχειρηματικές διαδικασίες τους και σε ορισμένες περιπτώσεις, την ενσωμάτωση αυτών των διαδικασιών ώστε να επικεντρώνονται στην αποτελεσματική διαχείριση των πόρων και την εξυπηρέτηση πελατών.

Σε κάθε επίπεδο της οργάνωσης τα πληροφοριακά συστήματα υποστηρίζουν τους κύριους λειτουργικούς τομείς της επιχείρησης. Οι πωλήσεις και τα συστήματα μάρκετινγκ βοηθούν την εταιρεία στην ταυτοποίηση των πελατών για τα προϊόντα ή τις υπηρεσίες της επιχείρησης, την ανάπτυξη προϊόντων και υπηρεσιών για την κάλυψη των αναγκών των πελατών, την προώθηση των προϊόντων και των υπηρεσιών, τη διαδικασία πώλησης των προϊόντων και υπηρεσιών, ενώ παρέχουν συνεχή υποστήριξη στον πελάτη.

Η κατασκευή και παραγωγή συστημάτων ασχολείται με το σχεδιασμό, την ανάπτυξη και την παραγωγή των προϊόντων ή των υπηρεσιών, καθώς και τον έλεγχο της ροής της παραγωγής. Επιπλέον, τα οικονομικά και λογιστικά συστήματα συμβάλλουν στην παρακολούθηση των οικονομικών και περιουσιακών στοιχείων και τις ροές κεφαλαίων της επιχείρησης. Κάτι εξίσου σημαντικό είναι ότι τα συστήματα ανθρώπινων πόρων διατηρούν αρχεία των εργαζομένων, το τμήμα δεξιοτήτων των εργαζομένων, την απόδοση στην εργασία και την κατάρτιση, καθώς και να στηρίξει το σχεδιασμό αποζημίωσης των εργαζομένων και την εξέλιξη της σταδιοδρομίας τους.

Η τεχνολογία της πληροφορίας και τα πληροφοριακά συστήματα μπορεί να είναι το “καμάρι” της εποχής και το σημείο μας στην πρόοδο του ανθρώπινου γένους, αλλά υπάρχουν αμέτρητα μειονεκτήματα για τα οποία συχνά οι άνθρωποι αναρωτιούνται αν το καλό υπερτερεί του κακού.

Μερικά μειονεκτήματα της τεχνολογίας των πληροφοριών περιλαμβάνουν τα εξής (Atzani, Bugiotti, and Rossi, 2012):

- Προβλήματα ανεργίας, πολλά παραδοσιακά επαγγέλματα χάνονται.
- Προβλήματα κοινωνικοποίησης των ατόμων στις σύγχρονες κοινωνίες.
- Αντιμετώπιση ζητημάτων ασφάλειας δεδομένων.
- Κοινωνικός αποκλεισμός, ΤΠΕ και πρόσβαση στην πληροφορία.

Τα πληροφοριακά συστήματα στην πορεία έχουν ενσωματωθεί μέσα από ειδικά προγράμματα όπως είναι τα ακόλουθα (Atzeni, Bugiotti, and Rossi, 2012):

1. Τα Decision Support Systems (DSS) είναι ένας υπολογιστής που βασίζεται σε σύστημα πληροφοριών. Υποστηρίζει τις διαδικασίες λήψης αποφάσεων σε επιχειρήσεις και οργανισμούς. Ένα ορθά δομημένο DSS είναι ένα διαδραστικό εργαλείο που βασίζεται σε λογισμικό σύστημα που

προορίζεται να βοηθήσει τους ιθύνοντες να συγκεντρώσουν χρήσιμες πληροφορίες από ένα συνδυασμό των πρωτογενών στοιχείων, εγγράφων, προσωπική γνώση, ή επιχειρηματικά μοντέλα για τον εντοπισμό και την επίλυση προβλημάτων και τη λήψη αποφάσεων.

2. Ο πρωταρχικός σκοπός των συστημάτων ERP είναι να διευθύνουν την επιχείρηση πολύ καλύτερα από ό, τι πριν, σε ένα ταχέως μεταβαλλόμενο και άκρως ανταγωνιστικό περιβάλλον. Σε μια επιχείρηση παραγωγής, η αλλαγή δεν είναι απλώς μια δυνατότητα αλλά μια πιθανότητα. Είναι μια βεβαιότητα, η μόνη σταθερά, το μόνο σίγουρο. Αυτό περιλαμβάνει υψηλά επίπεδα εξυπηρέτησης πελατών, την παραγωγικότητα, τη μείωση του κόστους και του κύκλου εργασιών των αποθεμάτων, καθώς και να παρέχει τη βάση για την αποτελεσματική διαχείριση της εφοδιαστικής αλυσίδας και του ηλεκτρονικού εμπορίου.

1.6 Τύποι πληροφοριακών συστημάτων

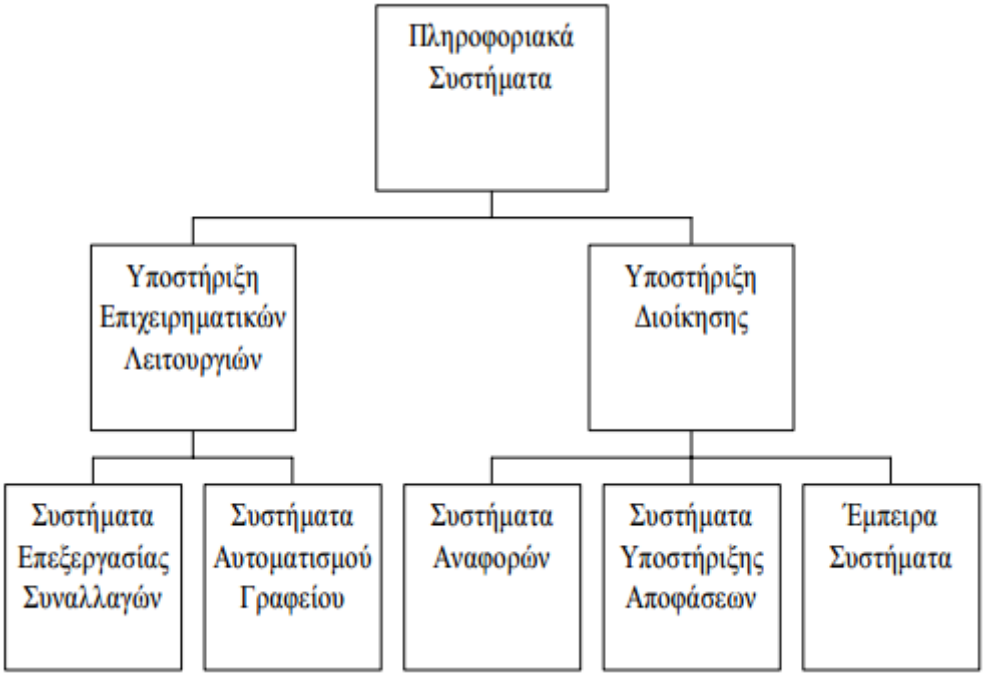
Υπάρχουν πολλοί τύποι πληροφοριακών συστημάτων στην αγορά. Τα επιχειρησιακά (ή οργανωτικά) πληροφοριακά συστήματα είναι μια σημαντική κατηγορία. Ένα πληροφοριακό σύστημα της επιχείρησης είναι προσαρμοσμένο ως προς την υποστήριξη εντός του οργανισμού και δύναται να κατηγοριοποιηθεί σύμφωνα με τα εξής: α) το υποσύστημα το οποίο υποστηρίζει, β) την επιχειρηματική δραστηριότητα που υποστηρίζει, γ) το είδος της υποστήριξης που παρέχει, δ) ανάλογα με την αρχιτεκτονική του. Τα επιχειρησιακά πληροφοριακά συστήματα υποστηρίζουν λειτουργίες που μπορεί να χρησιμοποιηθούν από ένα ευρύ φάσμα οργανισμών. Μερικά παραδείγματα είναι: τα συστήματα διαχείρισης ροής εργασιών, των επιχειρηματικών πόρων, τα συστήματα σχεδιασμού, τα συστήματα αποθήκευσης δεδομένων, τα γεωγραφικά συστήματα πληροφοριών κλπ.

Τα συστήματα αυτά δύναται να είναι είτε αυτόνομα ή συνδεδεμένα μεταξύ τους. Πληροφοριακά συστήματα σύμφωνα με την ιεραρχική δομή είναι (Paul, 2014):

1. Π.Σ. για τα τμήματα της επιχείρησης – πολλές φορές μία επιχείρηση αξιοποιεί ένα σύνολο εφαρμογών (προγράμματα) σε μία λειτουργική περιοχή. Οι εφαρμογές αυτές γενικά περιλαμβάνουν κοινά σημεία, ενώ άλλες φορές όχι. Το σύνολο των εφαρμογών που χρησιμοποιείται από το τμήμα προσωπικού αφορά το ίδιο τμήμα. Έχει σχεδιαστεί για την παρακολούθηση διατήρησης και ενημέρωσης των εργαζομένων και οτιδήποτε αφορά σε αυτούς.
2. Π.Σ. για όλη την επιχείρηση - τα Π.Σ. τα οποία περιλαμβάνουν όλα τα τμήματα της επιχείρησης και σχετίζονται με κάποια λειτουργία. Περιλαμβάνει έναν μεγάλο αριθμό εφαρμογών που υποστηρίζει όλες τις λειτουργίες της επιχείρησης.
3. Διεπιχειρησιακά Π.Σ. - αυτός ο τύπος Π.Σ. συγκεντρώνουν μεγάλο αριθμό οργανισμών. Για παράδειγμα, το παγκόσμιο σύστημα κράτησης θέσεων σε πτήσεις αποτελείται από τα συστήματα που ανήκουν σε διαφορετικές αεροπορικές εταιρίες.

Τα βασικότερα Π.Σ. είναι το λογιστικό, το οικονομικό, το Π.Σ. παραγωγής, το Π.Σ. προώθησης πωλήσεων και το Π.Σ. προσωπικού. Για το καθένα από αυτά παρατηρούνται ενέργειες καθημερινότητας οι οποίες

όμως είναι βασικές για την ορθή λειτουργία της επιχείρησης. Τα συστήματα σύμφωνα με αυτό το τρόπο κατηγοριοποίησης χωρίζονται σε δύο μεγάλες κατηγορίες (Βλέπε παρακάτω σχήμα 2) (Tsvetkov, 2014):
1.Π.Σ. που υποστηρίζουν τις λειτουργίες της επιχείρησης: Τα συγκεκριμένα αναφέρονται σε συστήματα επεξεργασίας συναλλαγών και σε συστήματα αυτοματοποίησης γραφείου.
2. Π.Σ. που υποστηρίζουν την διοίκηση: Τα συγκεκριμένα αναφέρονται σε αναφορές, συστήματα λήψης αποφάσεων και έμπειρα συστήματα.



Σχήμα 2. Τύποι Π.Σ. ανάλογα με το είδος της υποστήριξης που παρέχουν
Οι κύριες κατηγορίες είναι Π.Σ. που βασίζονται σε:

1. Βασικούς υπολογιστές: Η επεξεργασία υλοποιείται από ένα pc στο οποίο υπάρχουν συνδεδεμένα τερματικά δίχως υπολογιστική δυνατότητα (dump terminals). Η αρχιτεκτονική ήταν δημοφιλής τη δεκαετία του '80.
2. Προσωπικούς υπολογιστές – όπου αυτοί οι υπολογιστές δύναται να είναι συνδεδεμένοι μεταξύ τους. Η αρχιτεκτονική αυτή ενδείκνυται για μικρές και μεσαίες επιχειρήσεις.
3. Κατανεμημένα συστήματα - η επεξεργασία διαιρείται σε έναν αριθμό υπολογιστών οι οποίοι λειτουργούν σε διαφορετικά γεωγραφικά μήκη.

1.7 Ιστορική εξέλιξη των πληροφοριακών συστημάτων

Η πρώτη επιχειρησιακή εφαρμογή των ηλεκτρονικών υπολογιστών υλοποιήθηκε στα μέσα της δεκαετίας του 1950 όπου διενεργούσε επαναλαμβανόμενα, μεγάλου όγκου καθήκοντα των συναλλαγών-υπολογιστών. Έτσι οι υπολογιστές συνόψιζαν και την οργάνωση των συναλλαγών και τα δεδομένα της λογιστικής, τη χρηματοδότηση και τους ανθρώπινους πόρου. Τέτοια συστήματα γενικά ονομάζονταν συστήματα επεξεργασίας συναλλαγών.

Στη δεκαετία του 1960 ο ρόλος των πληροφοριακών συστημάτων ήταν απλός. Χρησιμοποιούνταν κυρίως για την ηλεκτρονική επεξεργασία δεδομένων με λειτουργίες όπως η επεξεργασία συναλλαγών, η τήρηση αρχείων και η λογιστική. Επίσης, ένας άλλος ρόλος προστέθηκε με τη χρήση των ηλεκτρονικών υπολογιστών: η επεξεργασία των δεδομένων σε χρήσιμες ενημερωτικές εκθέσεις. Η έννοια των συστημάτων διαχείρισης πληροφοριών (MIS) μόλις είχε γεννηθεί. Ο νέος αυτός ρόλος επικεντρωνόταν στην ανάπτυξη επιχειρηματικών εφαρμογών υπό την προϋπόθεση ότι περιελάμβανε διαχείριση και τελικούς χρήστες με προκαθορισμένες εκθέσεις διαχείρισης όπου αυτές οι πληροφορίες παρέχονται στους διευθυντές με τη μορφή εκθέσεων για να υποστηρίξουν την επιχειρησιακή λήψη αποφάσεων (πληροφοριακό σύστημα διοίκησης). Παραδείγματα αυτού: ανάλυση πωλήσεων, απόδοση παραγωγής και τάση δαπανών που εκθέτουν τα συστήματα (Xu, Chau, and Tan, 2014).

Στη δεκαετία του 1970 αυτές οι προκαθορισμένες εκθέσεις διαχείρισης δεν ήταν επαρκείς για να καλύψουν πολλές από τις ανάγκες της λήψης αποφάσεων της διοίκησης. Τότε υπολογιστικά συστήματα αξιοποίησαν το τηλεπικοινωνιακό δίκτυο (π.χ. συστήματα κράτησης θέσεων σε πτήσεις). Στην πορεία η χρήση τους διευρύνθηκε προκειμένου να ικανοποιηθούν καλύτερα οι ανάγκες των εταιρειών. Έτσι, άρχισαν να διαδίδεται η έννοια των συστημάτων υποστήριξης αποφάσεων. Ο νέος ρόλος για τα πληροφοριακά συστήματα ήταν η παροχή υποστήριξης των διαδικασιών λήψης των αποφάσεων στους διευθυντές και τελικούς χρήστες με έναν ad hoc και διαδραστικό τρόπο. Παραδείγματα αυτού του συστήματος ήταν: η τιμολόγηση των προϊόντων, η πρόβλεψη κερδοφορίας και τα συστήματα ανάλυσης κινδύνου.

Στη δεκαετία του 1980, η εισαγωγή των μικροϋπολογιστών στο χώρο εργασίας μπαίνει σε μια νέα εποχή, η οποία οδήγησε σε μια βαθιά επίδραση στις επιχειρήσεις. Η ταχεία ανάπτυξη της επεξεργασίας της μικροϋπολογιστής ισχύος (π.χ. Pentium μικροεπεξεργαστής της Intel), εφαρμογή πακέτα λογισμικού (π.χ. Microsoft Office), και τηλεπικοινωνιακά δίκτυα “γέννησε” τον τελικό χρήστη. Οι τελικοί χρήστες θα μπορούσαν πλέον να χρησιμοποιούν τους δικούς τους υπολογιστικούς πόρους για να υποστηρίξουν τις απαιτήσεις της εργασίας τους, αντί να αναμένουν την έμμεση υποστήριξη συγκεντρωμένη στις εταιρικές πληροφορίες του τμήματος υπηρεσιών. Επιπλέον, στα τέλη της δεκαετίας του '80 δημιουργήθηκαν τα συστήματα υποστήριξης ομάδων για την υποστήριξη των εργαζομένων σε ομάδες (Xu, Shau, and Tan, 2014).

1.8 Ασφάλεια πληροφοριακών συστημάτων και προβλήματα

Η ασφάλεια είναι ένας κλάδος της τεχνολογίας των υπολογιστών γνωστή ως η ασφάλεια των πληροφοριών η οποία εφαρμόζεται σε υπολογιστές και δίκτυα.

Ο στόχος της ασφάλειας στο διαδίκτυο περιλαμβάνει την προστασία των πληροφοριών και της περιουσίας από κλοπή, διαφθορά ή απειλές και επιθέσεις, επιτρέποντας παράλληλα την πληροφόρηση και τα δεδομένα να παραμένουν προσβάσιμα μόνο στους προβλεπόμενους χρήστες της επιχείρησης.

Παρά το γεγονός ότι η ασφάλεια των πληροφοριών παίζει σοβαρό ρόλο στην προστασία των δεδομένων και περιουσιακών στοιχείων κάθε επιχείρησης, δεν είναι λίγες οι φορές που ακούμε ειδήσεις για περιστατικά παραβίασης της ασφάλειας.
Αποτέλεσμα των παραβιάσεων και των επιθέσεων των λογισμικών των εταιρειών οδηγούν στην καταστροφή των πλαισίων όπως η εμπιστευτικότητα και η ακεραιότητα των πληροφοριών που διαχειρίζεται και συνεπώς ολοκληρωτική διάλυση της ασφάλειας των συστημάτων αυτών.

Ως εκ τούτου, οι σημερινές οργανώσεις δέχονται μεγαλύτερο κίνδυνο επιθέσεων των εμπιστευτικών πληροφοριών. Οι οργανισμοί πρέπει να υιοθετήσουν μια στρατηγική που να μπορεί να τους βοηθήσει να διαχειριστούν αποτελεσματικά τον κίνδυνο αυτό, ενώ θα πρέπει να είναι έτοιμοι να αποκρούσουν τις επιθέσεις οπουδήποτε και αν προέρχονται (Zafar, 2013).

1.8.1. Προϋποθέσεις ασφάλειας ενός Π.Σ.

Η ασφάλεια των πληροφοριακών συστημάτων θεωρείται σπουδαίας σημασίας καθώς βασίζεται σε τρεις τεχνικές ιδέες οι οποίες είναι κρίσιμες για την σωστή λειτουργία ενός Π.Σ., και είναι οι εξής (Sommestad, et al., 2011):

- Ακεραιότητα (Integrity): Εξασφάλιση ότι οι πληροφορίες και τα προγράμματα δεν θα αλλάξουν, αλλοιωθούν ή τροποποιηθούν σε ένα πληροφοριακό σύστημα αποτρέποντας από τρίτους να έχουν πρόσβαση και να μπορούν να τροποποιήσουν ούτε να διαμορφώσουν το πεδίο τους χωρίς να έχουν την απαιτούμενη άδεια. Για παράδειγμα, μια εφημερίδα η οποία δημοσιεύει τα άρθρα της και στο Διαδίκτυο επιθυμεί τα εν λόγω άρθρα να παραμείνουν ασφαλή από μεταβολές ενός χάκερ που επιδιώκει να εισάγει λανθασμένες πληροφορίες στα κείμενα.
- Διαθεσιμότητα (Availability): Διασφαλίζεται ότι οι εξουσιοδοτημένοι χρήστες έχουν συνεχή και έγκαιρη πρόσβαση στις πληροφορίες και τους πόρους - για παράδειγμα, εμποδίζοντας τον αντίπαλο από τις πλημμύρες ένα δίκτυο με ψεύτικο κυκλοφορίας που καθυστερεί νόμιμο κίνησης όπως αυτές που περιέχουν οι νέες παραγγελίες από το να μεταδίδονται. Μία συνηθισμένη απειλή που αντιμετωπίζουν τα σύγχρονα πληροφοριακά συστήματα είναι η επίθεση άρνησης υπηρεσιών (DOS attack), που έχει ως σκοπό να τεθούν εκτός λειτουργίας οι στοχευμένοι πόροι, είτε προσωρινά είτε μόνιμα.
- Εμπιστευτικότητα (Confidentiality): Η εμπιστευτικότητα σημαίνει ότι ο χρήστης μπορεί να ελέγχει ποιος μπορεί να παίρνει και να διαβάζει τις πληροφορίες προκειμένου να διατηρεί τις ευαίσθητες πληροφορίες ώστε να μην αποκαλύπτονται σε μη εξουσιοδοτημένα αποδέκτες -π.χ., την παρεμπόδιση της αποκάλυψης διαβαθμισμένες πληροφορίες σε έναν αντίπαλο ή με την κλοπή φορητών υπολογιστών από το συγκεκριμένο τμήμα μιας εταιρίας. Το 2006 μια μελέτη που υλοποιήθηκε με τη συμμετοχή 480 εταιριών φανέρωσε ότι 80% των εταιριών είχε πρόβλημα με διαρροή πληροφοριών λόγο κλοπής φορητού.

Σχήμα 3. Βασικές αρχές ασφάλειας

1.8.2. Εμπλεκόμενοι στην ανάπτυξη πολιτικών ασφαλείας

Η δημιουργία της πολιτικής ασφάλειας των πληροφοριακών συστημάτων μιας επιχείρησης στηρίζεται στην καταγραφή των απαιτήσεων ασφάλειας, με βάση τις οποίες οργανώνονται οι στόχοι της ασφάλειας, και στον σχεδιασμό των τρόπων για την επίτευξη των στόχων αυτών. Οι απαιτήσεις ασφάλειας δύναται να αφορούν διαφορετικές πηγές, όπως (Susanto, Almunawar, and Tuan, 2011):

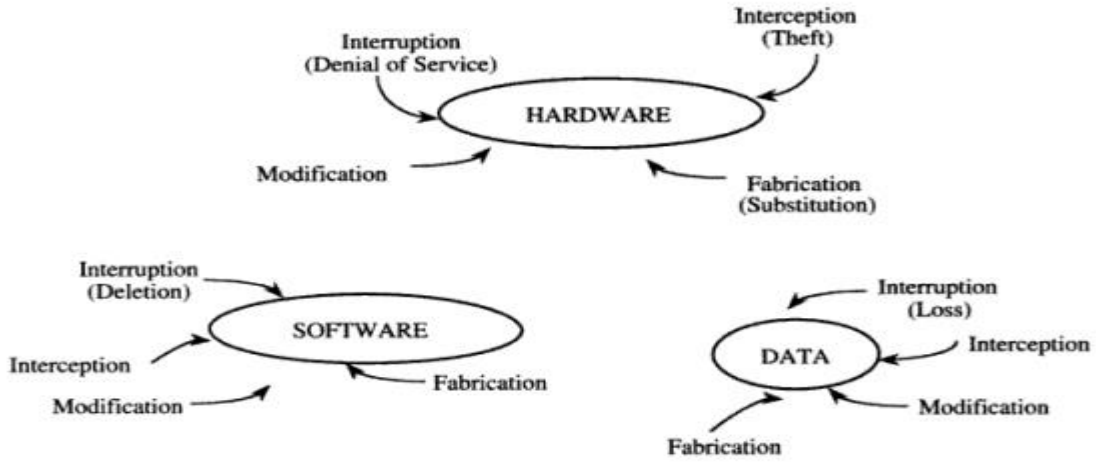
1. Οι χρήστες των πληροφοριακών συστημάτων.
2. Η διοίκηση του οργανισμού που επιθυμεί την απρόσκοπτη χρήση των πληροφοριακών συστημάτων στις λειτουργίες του οργανισμού.
3. Οι πελάτες του οργανισμού, εφόσον δεδομένα που τους αφορούν αποτελούν συνιστώσα του πληροφοριακού συστήματος.
4. Το νομικό και ρυθμιστικό πλαίσιο στο οποίο λειτουργεί ο οργανισμός.

Η πολιτική ασφάλειας θα πρέπει να καλύπτει όλες τις απαιτήσεις που παρουσιάζονται για τα πληροφοριακά συστήματα προκειμένου να εξασφαλιστεί το επιδιωκόμενο επίπεδο ασφάλειας.

1.8.3. Ανάλυση επικινδυνότητας

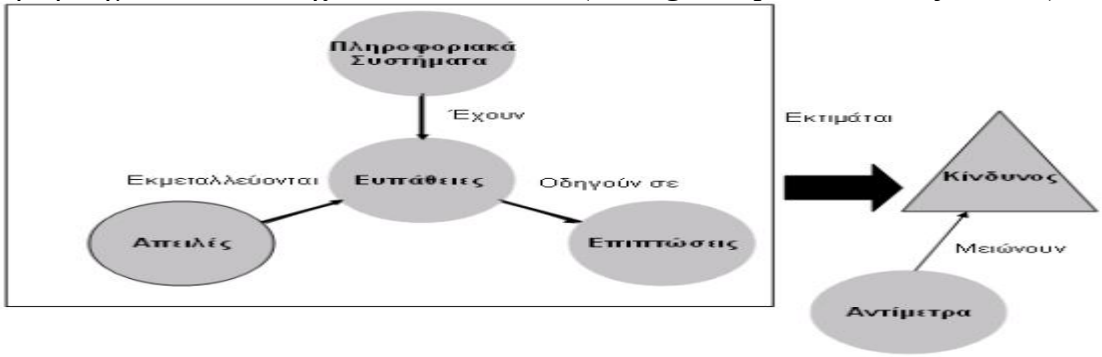
Ο καθορισμός των ακριβών απαιτήσεων ασφαλείας για μια συγκεκριμένη οργάνωση είναι απαραίτητη για την εφαρμογή των κατάλληλων μέτρων ασφαλείας. Τα μέτρα αυτά αποσκοπούν στην προστασία των συστημάτων πληροφοριών από παραβιάσεις και επιθέσεις. Πριν όμως προσδιοριστούν τα απαραίτητα μέτρα χρειάζεται να υπάρξει μια αξιολόγησή τους. Αυτό επιτυγχάνεται με διάφορους τρόπους, οι πιο συνηθισμένοι από αυτούς είναι η πραγματοποίηση μιας μελέτης ανάλυσης επικινδυνότητας (Risk Analysis) η οποία αποτελεί μια τυπική διαδικασία προσδιορισμού των κινδύνων και απειλών και η ανάπτυξη ενός σχεδίου για την αντιμετώπισή τους (Wang, Gupta, and Raj, 2015).

1. Απειλή: είναι μια δηλωμένη πρόθεση να επιβληθεί βλάβη ή ζημιά ή άλλο πρόβλημα σε ένα σύστημα και τις υπηρεσίες για την μετατροπή των δεδομένων, καταστροφή τους ή του συστήματος με αποτέλεσμα την πρόσβαση σε ευαίσθητες πληροφορίες.
2. Ευπάθεια: Ένα ελάττωμα ή αδυναμία στις διαδικασίες ασφάλειας του συστήματος κατά το σχεδιασμό, την υλοποίηση, ή τους εσωτερικούς ελέγχους που θα μπορούσαν να ασκηθούν (τυχαία ή σκόπιμα) και να οδηγήσει σε παραβίαση της ασφάλειας ή παραβίασης της πολιτικής ασφάλειας του συστήματος
3. Κίνδυνος: Ένας κίνδυνος είναι κανονικά ένα προϊόν δύο παραγόντων: απειλής (κάτι πάει στραβά) και ευπάθειας (τα πληροφοριακά συστήματα που χρησιμοποιούνται από την επιχείρηση για κάποιο λόγο επιτρέπουν στα πράγματα να λειτουργήσουν λανθασμένα)
4. Αντίμετρο: Όλα τα αντίμετρα ασφαλείας έχουν ως ευρύ στόχο την προσαρμογή της συμπεριφοράς των δυνητικών παραγόντων κινδύνου, έτσι ώστε να μην αποτελούν απειλή για τον οργανισμό



Σχήμα 1.4. Ευπάθειες ενός πληροφοριακού συστήματος

Για να είναι χρήσιμη μια διαδικασία ανάλυσης του κινδύνου θα πρέπει να παράγει μια ποσοτική δήλωση, τις επιπτώσεις ενός κινδύνου ή την επίδραση συγκεκριμένων προβλημάτων ασφάλειας. Τα τρία βασικά στοιχεία στην ανάλυση κινδύνου είναι: 1) η δήλωση επιπτώσεων ή το κόστος ενός συγκεκριμένου προβλήματος εφόσον συμβαίνει, 2) ένα μέτρο αποτελεσματικότητας των επί τόπου αντιμέτρων που θα υιοθετηθούν και 3) μια σειρά συστάσεων για τη διόρθωση ή την ελαχιστοποίηση των προβλημάτων που έχουν εντοπιστεί (Wang, Gupta, and Raj, 2015).



Σχήμα 1.5. Συσχέτισης των παραγόντων της ανάλυσης επικινδυνότητας

1.8.4 Μέτρα ασφαλείας

Η πολιτική ασφαλείας περιλαμβάνει πρότυπα (μέτρα ασφαλείας, μέτρα προστασίας, αντίμετρα) για την παροχή της ασφάλειας του συστήματος πληροφοριών που είναι απαραίτητο σε τέτοιες περιπτώσεις. Τα πρότυπα μπορούν να καθορίζουν το πεδίο εφαρμογής των λειτουργιών ασφαλείας και τα χαρακτηριστικά που απαιτούνται για την ορθή διαχείριση των πληροφοριών. Αυτό θα περιλαμβάνει όλες τις συσκευές που έχουν υποστεί τη βλάβη-απειλές του πληροφοριακού συστήματος, καθώς και το σχεδιάγραμμα πραγματοποίησής τους.

Τα αντίμετρα χωρίζονται σε 4 μεγάλες κατηγορίες (Qureshi, 2012):
α) Πρόληψη: Αυτή η κατηγορία έχει ως στόχο την πρόληψη της εμφάνισης των κινδύνων με τη χρήση ορισμένων μέτρων. Η πρόληψη του κινδύνου σημαίνει μείωση του κινδύνου μέχρι την αποδεκτή τιμή για την εκπλήρωση των στόχων των ενδιαφερόμενων μερών
β)Διασφάλιση: τεχνικές, εργαλεία, έρευνα και έλεγχοι οι οποίοι θα βοηθήσουν στην αποτελεσματικότητα και διατήρηση των υπαρχόντων αντιμέτρων
γ) Ανίχνευση: τα συστήματα ανίχνευσης αποσκοπούν στο να ανιχνεύσουν τις επιθέσεις αλλά και να χρησιμοποιηθούν μηχανισμοί, τεχνικές και προγράμματα για την αντιμετώπιση του προβλήματος
δ) Διόρθωση: διεργασίες που σκοπό έχουν στην γρήγορη επαναφορά σε ένα ασφαλές περιβάλλον καθώς και μελέτη των λόγων οι οποίοι προξένησαν τη ζημιά.

Τα μέτρα ασφαλείας και τα αντίμετρα χρησιμοποιούνται σε οργανώσεις για την προστασία και την ασφάλεια των πληροφοριακών συστημάτων. Η εφαρμογή αυτών των μέτρων θα βοηθήσουν στην διατήρηση ενός σημαντικού επιπέδου ασφάλειας. Για παράδειγμα, το επίπεδο ασφάλειας είναι υψηλό όταν ένας οργανισμός υλοποιεί τα πιο σωστά, ενημερωμένα μέτρα, τις πολιτικές και τα αντίμετρα για την υλοποίηση του στόχου που είναι η διασφάλιση της προστασίας των πληροφοριακών συστημάτων του οργανισμού (Qureshi, 2012).

Κεφάλαιο 2ο Εισαγωγή στην εξόρυξη δεδομένων

2.1 Ιστορία και Εξέλιξη

"Η χειροκίνητη εξαγωγή προτύπων από δεδομένα συμβαίνει εδώ και αιώνες. Οι πρώτοι μέθοδοι για τον προσδιορισμό προτύπων ήταν αυτοί της θεωρίας Bayes και της ανάλυσης της παλινδρόμησης. Ο πολλαπλασιασμός, η ευρεία διαθεσιμότητα και η εξέλιξη της τεχνολογίας υπολογιστών έχουν αυξήσει τον όγκο των συγκεντρωμένων δεδομένων και την ζήτηση για αποδοτικούς και αποτελεσματικούς χειρισμούς (Kantardzic, 2003)".

Καθώς οι συλλογές δεδομένων αυξήθηκαν τόσο σε όγκο όσο και σε πολυπλοκότητα, η χειρωνακτική ανάλυση των δεδομένων έχει αντικατασταθεί από την αυτόματη επεξεργασία δεδομένων.

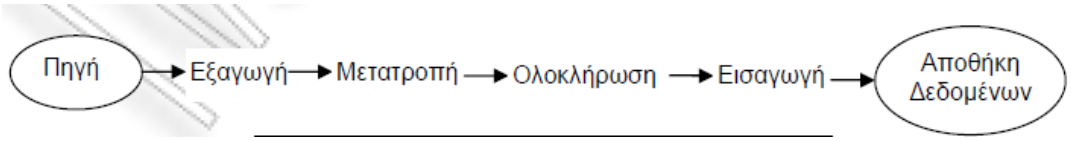
Σε αυτό συνέβαλαν άλλες ανακαλύψεις της επιστήμης των υπολογιστών, όπως τα νευρωνικά δίκτυα, η συσταδοποίηση, οι γενετικοί

αλγόριθμοι (1950), τα δέντρα απόφασης (1960) και η μηχανή υποστήριξης διανυσμάτων(1990). Η εξόρυξη δεδομένων είναι η διαδικασία εφαρμογής αυτών των μεθόδων στα δεδομένα με σκοπό την αποκάλυψη άγνωστων προτύπων σε μεγάλα σύνολα δεδομένων.

"Αυτό γεφυρώνει το χάσμα της εφαρμοσμένης στατιστικής και της τεχνητής νοημοσύνης (τα οποία συνήθως παρέχουν το μαθηματικό υπόβαθρο) με την διαχείριση βάσης δεδομένων κάνοντας χρήση του τρόπο με τον οποίο αποθηκεύονται και κατατάσσονται στη βάση δεδομένων για να εκτελέσουν την θεωρία και τους διαθέσιμους αλγορίθμους περισσότερο αποτελεσματικά, επιτρέποντας σε τέτοιες μεθόδους να εφαρμόζονται σε μεγάλα σύνολα δεδομένων" (Kantardzic, 2003).

2.2 Διαδικασία Εξόρυξης

Ο κύριος παράγοντας προκειμένου οι Αποθήκες Δεδομένων να είναι αποτελεσματικές είναι η σωστή τροφοδοσία που πραγματοποιείται από τις πηγές. Η διακομιδή των δεδομένων αυτών στην Αποθήκη Δεδομένων είναι περίπλοκη για το λόγο ότι υπάρχουν προβλήματα τα οποία είναι αναγκαίο να επιλυθούν. Στο Σχήμα 1 υπάρχει η διαδικασία που ακολουθείται και αφορά την διακομιδή των δεδομένων.



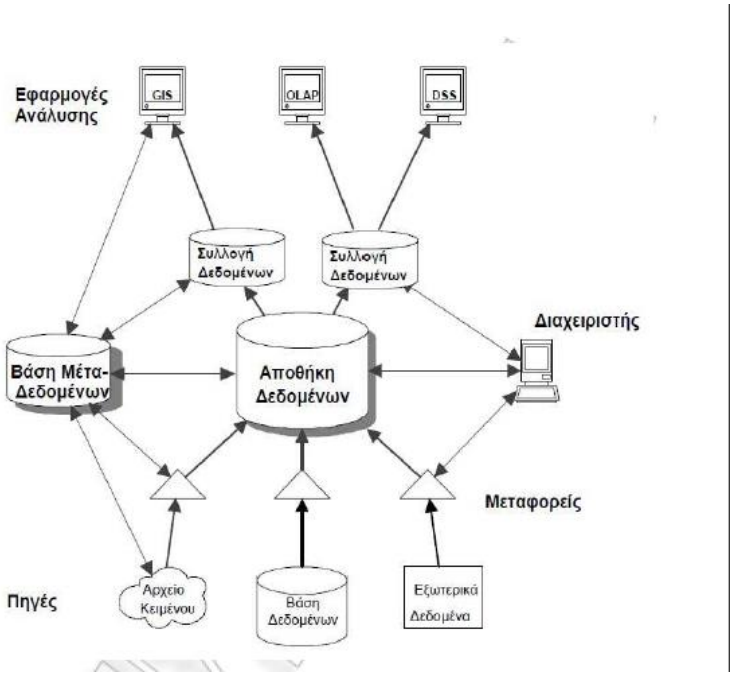
Σχήμα 1 : Διαδικασία μεταφοράς δεδομένων

2.2.1 Εξαγωγή και Μετατροπή Δεδομένων

Όπως φαίνεται στο Σχήμα 2, η Εξαγωγή και η Μετατροπή δεδομένων, πραγματοποιείται από τους Μετατροπείς / Μεταφορείς. Η εγκατάσταση ενός λογισμικού σε κάθε ξεχωριστή πηγή, βοηθάει στην άντληση δεδομένων, τα οποία τα φιλτράρει και χρησιμοποιεί μόνο αυτά που της είναι χρήσιμα, βάσει ενός προτύπου. Όσες μετατροπές γίνονται έχουν σχέση με την τιμή και τη δομή τους. Με άλλα λόγια, σε έναν πίνακα η κατηγορία “Ημερομηνία” είναι δυνατόν να αναφερθεί ως “Χρόνος”, “Μήνας” ή “Ημέρα” και η κατηγορία “Χαρακτηρισμός” ως “κατηγορία Α ή Β”. Το συγκεκριμένο λογισμικό περιλαμβάνει τα γνωρίσματα της πηγής και υπάρχει σε υπολογιστές, οι οποίοι έχουν απευθείας πρόσβαση στα δεδομένα της κάθε πηγής. Είναι γεγονός ότι υπάρχουν πολλά εργαλεία που εξυπηρετούν στην εξαγωγή δεδομένων. Οι πηγές που βρίσκονται μακριά χρησιμοποιούν τις πύλες (gateways) και συγκεκριμένες εφαρμογές σύνδεσης, όπως είναι για παράδειγμα η ODBC ή Oracle Open Connect για την εξαγωγή δεδομένων.

Επίσης χρησιμοποιούνται εξωτερικά εργαλεία, τα οποία βοηθούν στην εξαγωγή των δεδομένων. Την ίδια στιγμή επεξεργάζονται για πρώτη φορά τα δεδομένα αυτά. Εφόσον μέσω των Αποθηκών Δεδομένων συμπεραίνονται στρατηγικές αποφάσεις θα πρέπει να υπάρχουν ορθά δεδομένα. Όποια πηγή έχει πολλά δεδομένα ίσως αντιμετωπίσει προβλήματα ή λάθη.

Ακόμη, υπάρχουν εργαλεία μέσω των οποίων είναι δυνατόν να προσδιοριστούν και να επιλυθούν τα λάθη των δεδομένων, όπου βέβαια αυτό μπορεί να γίνει. Ο καθορισμός των δεδομένων είναι απαραίτητος στις εξής περιπτώσεις: α) όταν υπάρχει ασυνέπεια στο μήκος των κατηγοριών ανόμοιων πηγών β) όταν υπάρχει ασυνέπεια στην περιγραφή των δεδομένων, γ) όταν υπάρχει ασυνέπεια στις τιμές των δεδομένων δ) όταν δεν υπάρχουν εγγραφές και παραβιάζονται οι διάφοροι περιορισμοί.



Σχήμα 2 Γενική Αρχιτεκτονική Αποθήκης Δεδομένων

2.3 Ολοκλήρωση

Η ολοκλήρωση (integration) των Δεδομένων είναι περίπλοκη και έχει να κάνει με τη δημιουργία και τη διατήρηση του ιδεατού σχήματος των Δεδομένων. Το συγκεκριμένο σχήμα περικλείει τα δεδομένα που δίνει κάθε οντότητα ανεξάρτητα από την πηγή δεδομένων. Ολόκληρο το σχήμα διατηρείται χάρη στη σωστή πληροφόρηση της βάσης των μετα-δεδομένων. Κάθε όγκος δεδομένων που προέρχεται από τις πηγές, βάσει του σχήματος είναι δυνατόν να αλλάξει προκειμένου να ενταχθεί στην Αποθήκη Δεδομένων.

Με άλλα λόγια, σε έναν οργανισμό τηλεπικοινωνίας μπορούν δύο συστήματα να διευθετούν χρεώσεις για δύο εντελώς διαφορετικούς καταναλωτές. Στα δύο αυτά πληροφοριακά συστήματα ο πίνακας που αφορά την “Χρέωση” έχει άλλη έννοια στο ένα σύστημα και άλλη στο άλλο. Αφού ολοκληρωθούν οι πηγές σε όλο το σχήμα, η “Χρέωση”, ορίζεται από τη μία από τους επιμέρους ορισμούς, από την άλλη είναι δυνατόν να μην ορίζεται ούτε στον έναν ούτε στον άλλον. Οι αντίστοιχες εγγραφές των χρεώσεων που πραγματοποιούνται στα δύο συστήματα είναι αναγκαίο να αλλάξουν ώστε να γίνει η εισαγωγή τους στην Αποθήκη Δεδομένων.

2.4 Εισαγωγή δεδομένων

Η Εισαγωγή των Δεδομένων είναι η τελευταία διαδικασία σχετικά με τη διακομιδή των Δεδομένων προς την Αποθήκη. Μέσα από τις πράξεις αυτές θα βγουν τα Δεδομένα, τα οποία καταχωρούνται στην Αποθήκη και ταυτόχρονα γίνεται πληροφόρηση των ευρετηρίων της βάσης της Αποθήκης. Όσο πραγματοποιείται η εισαγωγή των Δεδομένων, είναι σκόπιμο αυτός που διαχειρίζεται τα Δεδομένα να επιβλέπει και να παρεμβαίνει όποτε είναι αναγκαίο. Αν και η Εισαγωγή Δεδομένων κοστίζει ακριβά, τόσο στις πηγές όσο και στην Αποθήκη, πραγματώνεται σε τακτά χρονικά διαστήματα.

2.4.1 Ενημέρωση

Η διαδικασία ενημέρωσης έχει να κάνει με τις μετατροπές των Δεδομένων των Πηγών και της Αποθήκης. Η συγκεκριμένη διαδικασία κάνει όλες τις παραπάνω ενέργειες δηλαδή, την Εξαγωγή, τη Μετατροπή, την Ολοκλήρωση, την Εισαγωγή. Παρόλα αυτά υπάρχουν ορισμένα προβλήματα που οφείλονται στις μετατροπές που γίνονται στις πηγές, όπως και στο μέγεθος των Δεδομένων που αλλάζουν. Τις περισσότερες φορές, οι Αποθήκες Δεδομένων πληροφορούνται σε τακτά χρονικά διαστήματα. Παρόλα αυτά ορισμένες εφαρμογές ανάλυσης χρειάζονται άμεσα τα δεδομένα. Έτσι είναι αναγκαίο οι Αποθήκες των Δεδομένων να ενημερώνονται για οποιαδήποτε αλλαγή που πραγματοποιείται στις πηγές. Η ενημέρωση, προσδιορίζεται από τον Διαχειριστή, ο οποίος στηρίζεται στις εφαρμογές ανάλυσης, τις πηγές που υπάρχουν και το δίκτυο που ενώνει τις πηγές με την Αποθήκη.

Η Ενημέρωση επηρεάζεται από τις Πηγές. Συχνά, η Εξαγωγή αφορά μόνο ένα ολόκληρο αρχείο ή μια Βάση Δεδομένων. Στη συγκεκριμένη περίσταση, η Ενημέρωση αντιστοιχεί με τη διαγραφή των Δεδομένων της Πηγής. Επίσης εισάγονται από την αρχή τα δεδομένα που συμπεράθηκαν. Η λύση αυτή δεν είναι και η σωστή, όμως συχνά είναι η μόνη, σε περιπτώσεις που η πηγή δεν μπορεί να δώσει πληροφορίες για τις αλλαγές που δέχεται.

Λαμβάνοντας υπόψη ότι μεγάλος όγκος Δεδομένων υπάρχει στις Αποθήκες Δεδομένων, η πιο πάνω Ενημέρωση δεν μπορεί να πραγματοποιηθεί. Για αυτό το λόγο θα πρέπει να προσδιοριστούν οι αλλαγές που γίνονται στις πηγές, δηλαδή οι εισαγωγές, οι διαγραφές και οι αλλαγές των εγγράφων, προκειμένου όταν πραγματοποιείται Ενημέρωση να μην υπάρχουν αναγκαίες διαγραφές και εισαγωγές Δεδομένων, τα οποία δεν έχουν δεχτεί κάποια αλλαγή. Κατά την προοδευτική Ενημέρωση και διατήρηση των Δεδομένων, οι Αποθήκες δέχονται νέες εγγραφές, οι οποίες απορρέουν από την εισαγωγή Δεδομένων σε Πηγές. Το ίδιο πράγμα συμβαίνει και όταν υπάρχουν διαγραφές και τροποποιήσεις εγγράφων, οι οποίες πάλι απορρέουν από τις ανάλογες πράξεις των Δεδομένων.

Προκειμένου να επιτευχθεί η προοδευτική Ενημέρωση είναι αναγκαίο οι Πηγές να μπορούν να προσδιορίζουν τις αλλαγές που γίνονται στα Δεδομένα τους. Στις περιστάσεις όπου μία εκσυγχρονισμένη βάση Δεδομένων είναι η Πηγή, ακολουθούνται τρεις τεχνικές οι οποίες επιτυγχάνουν τον προσδιορισμό των αλλαγών αυτών:

1. Στιγμιότυπα: Πολλές βάσεις Δεδομένων μπορούν να δώσουν, όταν είναι αναγκαίο στιγμιότυπα (snapshots), σύμφωνα με τους πίνακες που είναι στη βάση τους. Βάσει αυτών των στιγμιότυπων είναι δυνατόν να προσδιοριστούν οι αλλαγές που έγιναν στην Πηγή και η Αποθήκη να έχει τη σχετική Ενημέρωση.
2. Μηχανισμός καταγραφής (log): Οι πιο πολλές βάσεις δεδομένων καταχωρούν όλες τις αλλαγές των δεδομένων τους και τους λόγους που τις προκαλούν, με σκοπό να υπάρχει φυσιολογική εξέλιξη της δοσοληψίας. Εάν υπάρχει η δυνατότητα πρόσβασης, των Μεταφορέων, στο αρχείο (log file), όπου υπάρχουν όλες αυτές οι αλλαγές, τότε είναι δυνατόν να έχουν πρόσβαση άμεσα στις αλλαγές που γίνονται.
3. Triggers: Εάν μία πηγή είναι ένα εκσυγχρονισμένο σύστημα, μέσω του οποίου είναι δυνατόν να δημιουργηθούν triggers, τότε υπάρχει η δυνατότητα για κάθε πίνακα Πηγής να δημιουργείται ένας trigger, το οποίο θα πληροφορεί για κάθε αλλαγή που πραγματοποιείται στον συγκεκριμένο πίνακα.

2.4.2 Η διαδικασία για την εξαγωγή process patterns

Βήμα 1 Μεθοδολογίες Ενοποίησης

Τα υφιστάμενα SDMs (Scientific data management systems) είναι μάλλον άκαμπτης προέλευσης, και δεν δημιουργήθηκαν για να είναι modular. Σε γενικές γραμμές, αντιπροσωπεύονται με λέξεις σε φυσική γλώσσα, έτσι ώστε να είναι πολύ δύσκολο να υποβληθούν σε επεξεργασία από υπολογιστές. Επιπλέον, το modularity τους είναι περιορισμένο σε τέτοιο βαθμό, ώστε να παρέχουν διάφορα μοντέλα και τις συναφείς δεσμευτικές κατευθυντήριες γραμμές για να κατασκευάσουν διαφορετικές πλευρές σε εφαρμογές λογισμικού. Ως εκ τούτου, το πρώτο βήμα είναι να εκπροσωπηθούν τα SDMS σε ομοιόμορφη δομή.

Η στρατηγική "διαδικασία με επίκεντρο το πρότυπο" θα επιτρέψει να υπάρξει μια ενιαία δομή των μεθοδολογιών έτσι ώστε η αναλυτική σύγκριση να γίνει εύκολη. Το πρότυπο αυτό χρησιμοποιείται για την ανάδειξη των δραστηριοτήτων που προβλέπονται σε κάθε SDM, διατηρώντας παράλληλα τις λεπτομέρειες του προϊόντος ως δευτερεύουσες για τις δραστηριότητες.

Η περιγραφή που παράγεται χρησιμοποιώντας αυτό το πρότυπο επιτρέπει λεπτομερέστερη ανάλυση των επιμέρους SDM προκειμένου να ανακαλύψουν επαναλαμβανόμενες δραστηριότητες που οδηγούν στην αναγνώριση προτύπων διαδικασίας. Η δομή ενός SDM που βασίζεται σε αυτό το πρότυπο περιγράφεται στον Πίνακα 1. Το αποτέλεσμα αυτής της στρατηγικής είναι το SDM να είναι ίσο με την προέλευσή του με ενιαίες και συγκρίσιμες δραστηριότητες.

Πίνακας 1 - Διαδικασία Προτύπου για Περιγραφή SDMS

Επισκόπηση	**Μια σύντομη εισαγωγική του SDM που διακρίνει έντονα χαρακτηριστικά, δυνατά σημεία, αδυναμίες και μια οπτική διαδικασία ανάπτυξης που περιγράφουν το SDM.**	
Περιγραφή SDM	Φάσεις του SDM	Υψηλού επιπέδου υπο-διεργασίες στη διαδικασία του SDM αποτελούνται από δραστηριότητες, σειρά με την οποία πραγματοποιούνται και συνοπτική περιγραφή των παραγόμενων προϊόντων εργασίας.
	Λεπτομέρειες εσωτερικών δραστηριοτήτων	Κάθε δραστηριότητα περιέχει ένα ή περισσότερα στάδια που περιγράφουν λεπτομέρειες από αυτά. Οι σχετικές δραστηριότητες τοποθετούνται σε χωριστές φάσεις του SDM

Βήμα 2: Προσδιορισμός Φάσης Patterns

Για να επιτευχθούν τα πρότυπα της διαδικασίας, η στρατηγική «πρότυπα επιλογής Φάσης» χρησιμοποιείται. Τα πρότυπα διαδικασίας Φάσης, στην πραγματικότητα αντιπροσωπεύουν τις γενικές φάσεις της ανάπτυξης Κύκλου Ζωής λογισμικού (Software Development Life Cycle - SDLC). Θα πρέπει να σημειωθεί ότι οι δραστηριότητες «ομπρέλας» έχουν αποκλειστεί από τον ορισμό. Αν και οι λεπτομέρειες των δραστηριοτήτων SDMS τα κάνουν να ξεχωρίζουν, σε επίπεδο φάσης δεν έχουν καμία σημαντική και καινοτόμο διαφορά. Ως εκ τούτου, η στρατηγική καθορίζει τη γενική διαδικασία της φάσης προτύπων, καθώς και τις φάσεις SDLC του. Θα πρέπει να σημειωθεί, σε ορισμένες περιπτώσεις, όταν έχει επιλεγεί ένας τομέας του SDMS ότι στην διαδικασία εξαγωγής process patterns, οι φάσεις θα πρέπει να θεωρούνται φάσεις process patterns αντί για SDLC.

Σε άλλες περιπτώσεις, οι SDLC φάσεις θεωρούνται φάσεις process patterns. Ενώ η πρόθεση των process patterns είναι απλή, ένα μέρος του επιλεγμένου προτύπου για την εκπροσώπησή τους έχει ολοκληρωθεί. Για παράδειγμα, ο πίνακας 2 δείχνει μια τυπική αναπαράσταση του προτύπου φάσης με βάση το προτεινόμενο πρότυπο. Τα πρότυπα process patterns

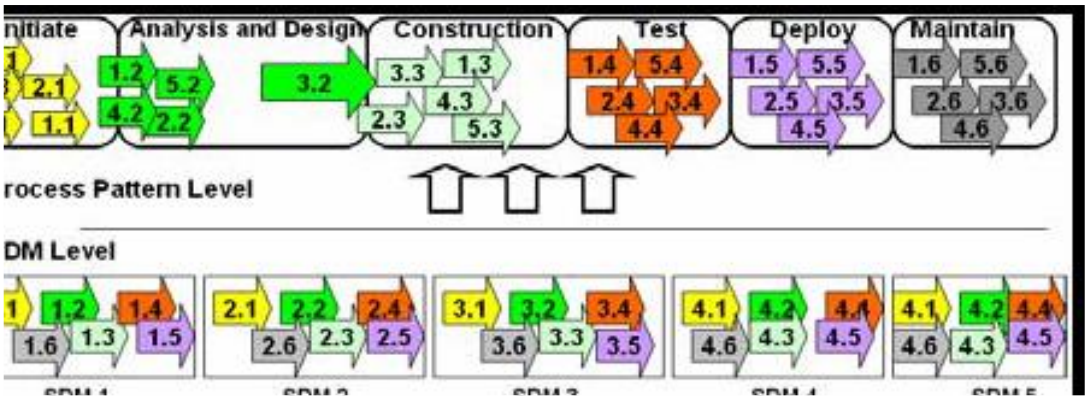
λειτουργούν ως πλαίσια για την κατηγοριοποίηση εσωτερικών δραστηριοτήτων του SDMS και θα χρησιμοποιηθούν στα ακόλουθα στάδια.

Πίνακας 2 – Αναπαράσταση δοκιμαστικής φάσης Pattern

Στοιχείο	**Περιγραφή**
Περιεχόμενο	Μια σειρά από ευρήματα έχουν παραχθεί και είναι έτοιμα να αξιολογήσουν πόσες απαιτήσεις και κριτήρια ποιότητας έχουν ικανοποιήσει.
Πρόβλημα	Πως μπορούν να αξιολογηθούν τα ευρήματα που παρήχθησαν
Διαδικασία pattern	Όλα τα στάδια και task patterns που περιλαμβάνονται στο pattern
Ρόλοι	Μηχανικός δοκιμών, δοκιμή script writer, δοκιμή executer.
Εύρημα	Test scripts, test results.
Σχετικά patterns	Αυτό το pattern αντιστοιχεί σε όλα τα πρότυπα φάσης.
Συνέπεια	Προς διερεύνηση

Βήμα 3: Αποδόμηση των SDMs

Η πρόθεση της «Αποδόμησης SDMS» είναι να αποκτήσουν ένα πλαίσιο για την ανάλυση των δραστηριοτήτων των SDMS. Για να επιτευχθεί αυτό, η «στρατηγική αποδόμησης» αποσυνθέτει τις δραστηριότητες των SDMS. Κάθε δραστηριότητα στην υποκείμενη SDM είναι υποψήφια για να οριστεί ως process pattern και πιο συγκεκριμένα ως task pattern. Αφού η «στρατηγική αποδόμησης» αποσυνθέτει τα SDMS βάζει τις εσωτερικές δραστηριότητες για τα αντίστοιχα πρότυπα της διαδικασίας φάση. Όπως δείχνει το σχήμα 3, οι διαφορετικές δραστηριότητες στις SDMS με το ίδιο χρώμα έχουν την ίδια πρόθεση και, ως εκ τούτου εμπίπτουν στο ίδιο μοτίβο φάσης.



Σχήμα 3- Δραστηριότητες αποδόμησης SDMS σε φάση process patterns.

Οι δραστηριότητες μπορούν να τοποθετηθούν εξ' ολοκλήρου σε μια φάση ή να μεσολαβήσουν μεταξύ δύο εξ αυτών. Σύμφωνα με τον αλγόριθμο 1 (Σχήμα 4), κάθε SDM περνάει σε process patterns. Ο χειριστής SYSNOMYM, ως αναλυτής ομοιότητας, ελέγχει κατά πόσον η φάση δραστηριότητας είναι συνώνυμη με ένα ή περισσότερα μοτίβα διαδικασίας φάσης.

Algorithm for Decomposing Methodologies
foreach *PhaseProcessPattern* P_i
  foreach *Methodology* M_j
    foreach $activity_k$ *in* M_j
      if ($activity_k.phase.name$ *SYNONYM* $P_i.phasename$)
        &&
        ($activity_k.phase.intent$ *SEMANTIC AFFINITY* $P_i.intent$) then
        $P_i \leftarrow P_i \cup activity_k$
    endfor
  endfor
endfor

. Σχήμα 4- Αλγόριθμος για την αποσύνθεση SDMS σε φάση process patterns

2.5 Τεχνικές

Το θέμα του Web Data Extraction (εξαγωγή δεδομένων από τον παγκόσμιο ιστό) έχει καλυφθεί από διάφορες κριτικές εκθέσεις. Οι Laender et al. παρουσίασαν μία έρευνα η οποία προσέφερε μία αυστηρή ταξινόμηση για την κατηγοριοποίηση των συστημάτων Web Data Extraction (Laender, et al., 2002).

Οι συγγραφείς εισήγαγαν μία σειρά κριτηρίων και μία ποιοτική ανάλυση των διαφόρων εργαλείων εξαγωγής δεδομένων από το διαδίκτυο. Ο Kushmerick καθόρισε ένα προφίλ όλων των προσεγγίσεων στο πρόβλημα του Web Data Extraction (Kushmerick, 2002). Ο συγγραφέας ανέλυσε τις προσεγγίσεις επαγωγής wrapper (δηλαδή, τις προσεγγίσεις

που είναι ικανές να παράγουν αυτόματα wrapper με την αξιοποίηση κατάλληλων παραδειγμάτων) και το πρόβλημα της συντήρησης wrapper (δηλαδή, τις μεθόδους για την επικαιροποίηση του wrapper κάθε φορά που η δομή της πηγής του ιστού αλλάζει).

Στην εν λόγω εργασία, συζητήθηκαν επίσης οι τεχνικές Web Data Extraction που προέρχονται από τα μοντέλα Natural Language Processing και Hidden Markov. Για το πρόβλημα της επαγωγής wrapper, ο Eikvil και οι Zung et al., ερεύνησαν διάφορες προσεγγίσεις, τεχνικές και εργαλεία. Η τελευταία αυτή έρευνα μάλιστα, παρουσίασε ένα μοντέλο που περιγράφει την αρχιτεκτονική ενός συστήματος Εξαγωγής Πληροφορίας (Eikvil, 1999; Jung, In Kim, and K Jain, 2004). Οι Chang et al. εισήγαγαν μία τρισδιάστατη κατηγοριοποίηση των συστημάτων Web Data Extraction, η οποία βασίζεται στις δυσκολίες έργου, τις τεχνικές που χρησιμοποιούνται και τον βαθμό αυτοματοποίησης (Chang, et al., 2006).

Το 2007, ο Fiumara εφήρμοσε αυτά τα κριτήρια για την ταξινόμηση των τεσσάρων συστημάτων Web Data Extraction τελευταίας τεχνολογίας (Fiumara, 2007). Μία σχετική έρευνα για την Εξαγωγή Πληροφορίας πραγματοποιήθηκε από τον Sarawagi (2008). Πρόσφατα, ορισμένοι συγγραφείς επικεντρώθηκαν στα συστήματα διαχείρισης μη δομημένων δεδομένων (UDMSs) , δηλαδή, τα συστήματα λογισμικού που αναλύουν τα ανεπεξέργαστα δεδομένα κειμένου, από τα οποία αποσπούν κάποια δομή (π.χ. το όνομα του ατόμου και την θέση), ενσωματώνουν την δομή (π.χ., αντικείμενα όπως η Νέα Υόρκη και NYC συγχωνεύονται σε ένα ενιαίο αντικείμενο) και χρησιμοποιούν την ολοκληρωμένη δομή για να χτίσουν μία βάση δεδομένων (Doan, et al., 2009). Τα UDMS αποτελούν ένα σχετικό παράδειγμα των συστημάτων Web Data Extraction και η εργασία των Doan et al. παρέχει μία επισκόπηση του Cimple, ενός UDMS το οποίο αναπτύχθηκε στο Πανεπιστήμιο του Wisconsin (Doan, et al., 2009). Η έρευνα των Baumgartner et al. αποτελεί μέχρι στιγμής την πιο πρόσφατα ενημερωμένη κριτική έκθεση σχετικά με το εξεταζόμενο θέμα (Baumgartner, et al., 2009).

Οι πρώτες προσπάθειες για την εξαγωγή δεδομένων από το διαδίκτυο χρονολογούνται στις αρχές της δεκαετίας του 1990, η οποία δανείστηκε προσεγγίσεις και τεχνικές από την βιβλιογραφία της εξαγωγής πληροφορίας (IE). Ειδικότερα, προέκυψαν δύο κατηγορίες στρατηγικών (Eikvil, 1999): εκμάθηση τεχνικών και τεχνικές γνώσεις μηχανικής, οι οποίες αποκαλούνται επίσης και προσεγγίσεις με βάση την μάθηση και προσεγγίσεις με βάση τους κανόνες, αντίστοιχα (Sarawagi, 2008). Αυτές οι κατηγορίες μοιράζονται μία κοινή λογική: η πρώτη θεωρείται ότι αναπτύσσει συστήματα που απαιτούν την ανθρώπινη πείρα για τον καθορισμό των κανόνων (π.χ. οι συνηθισμένες εκφράσεις) για να ολοκληρώσει με επιτυχία την εξαγωγή δεδομένων.

Αυτές οι προσεγγίσεις απαιτούν τεχνογνωσία σε συγκεκριμένο τομέα: οι χρήστες που σχεδιάζουν και εφαρμόζουν τους κανόνες και εκπαιδεύουν το σύστημα πρέπει να διαθέτουν εμπειρία στον προγραμματισμό και καλή γνώση του τομέα στον οποίο θα λειτουργήσει το σύστημα εξαγωγής δεδομένων. Θα πρέπει επίσης να έχουν την δυνατότητα να προβλέψουν πιθανά σενάρια χρήσης και καθήκοντα που θα

έχουν ανατεθεί στο σύστημα. Από την άλλη πλευρά, επίσης, κάποιες προσεγγίσεις της τελευταίας κατηγορίας συνεπάγονται μεγάλη εξοικείωση τόσο με τις απαιτήσεις, όσο και με τις λειτουργίες της πλατφόρμας, έτσι ώστε η εμπλοκή του ανθρώπου να είναι απαραίτητη.

Διάφορες στρατηγικές έχουν επινοηθεί για να μειώσουν την εμπλοκή του ανθρώπινου παράγοντα σε εξειδικευμένα θέματα πεδίου. Μερικές από αυτές έχουν αναπτυχθεί στο πλαίσιο της βιβλιογραφίας της Τεχνητής Νοημοσύνης, όπως η υιοθέτηση συγκεκριμένων αλγορίθμων που χρησιμοποιούν τη δομή των ιστοσελίδων για τον εντοπισμό και την εξαγωγή δεδομένων. Άλλες μέθοδοι προήλθαν από την Μηχανική Μάθηση, όπως οι εποπτευόμενες ή ημι-εποπτευόμενες τεχνικές μάθησης για τον σχεδιασμό συστημάτων ικανών να εκπαιδεύονται από παραδείγματα και στη συνέχεια να σε θέση να εξάγουν δεδομένα αυτόνομα από παρόμοια (ή ακόμα και διαφορετικά) πεδία.

2.6 Τεχνικές με βάση το δέντρο

Ένα από τα χαρακτηριστικά που έχει αξιοποιηθεί περισσότερο στο Web Data Extraction είναι η ημι-δομημένη φύση των ιστοσελίδων. Αυτές μπορούν φυσικά να παρασταθούν ως δέντρα με ρίζες σε διάταξη και με επισήμανση, όπου η επισήμανση αντιπροσωπεύει την ετικέτα που είναι σωστή για την σύνταξη της γλώσσας HTML mark-up, και η ιεραρχία του δέντρο αντιπροσωπεύει τα διαφορετικά επίπεδα της ενσωμάτωσης των στοιχείων που αποτελούν την ιστοσελίδα. Η αναπαράσταση μιας ιστοσελίδας με τη χρήση ενός επισημασμένου διατεταγμένου δέντρου με ρίζες συνήθως αναφέρεται ως DOM (Document Object Model).

Η γενική ιδέα πίσω από το Μοντέλο Αντικειμένου Εγγράφου (DOM) είναι ότι οι ιστοσελίδες HTML αναπαριστώνται μέσω του απλού κειμένου που περιέχει τις ετικέτες HTML, συγκεκριμένες λέξεις-κλειδιά που ορίζονται στην mark-up γλώσσα, η οποία μπορεί να ερμηνευθεί από το πρόγραμμα περιήγησης για να αναπαραστήσει τα ειδικά στοιχεία της ιστοσελίδας (π.χ. υπερ-συνδέσμους, κουμπιά, εικόνες και ούτω καθεξής), ως ελεύθερο κείμενο. Οι ετικέτες HTML μπορεί να είναι ένθετες η μία στην άλλη, σχηματίζοντας μια ιεραρχική δομή. Αυτή η ιεραρχία συλλαμβάνεται στο DOM από το δέντρο εγγράφου, του οποίου οι κόμβοι αναπαριστούν τις ετικέτες HTML. Το δέντρο εγγράφου (στο εξής δέντρο DOM) έχει αξιοποιηθεί με επιτυχία για τους σκοπούς του Web Data Extraction σε διάφορες τεχνικές.

2.7 Στοιχεία στο δέντρο εγγράφου: XPath

Ένα από τα κύρια πλεονεκτήματα της υιοθέτησης του Document Object Model για τη γλώσσα HTML είναι η δυνατότητα εκμετάλλευσης ορισμένων τυπικών εργαλείων των γλωσσών XML (και η HTML είναι μία διάλεκτος της XML). Ειδικότερα, η XML Path Language (ή, Briey, XPath) παρέχει μία ισχυρή σύνταξη για την αντιμετώπιση συγκεκριμένων στοιχείων ενός εγγράφου XML (και, στον ίδιο βαθμό και ιστοσελίδων HTML) με απλό τρόπο.

Υπάρχουν δύο πιθανοί τρόποι χρήσης της XPath: (α) προσδιορισμός ενός στοιχείου στο δέντρο εγγράφου, ή (β) για την αντιμετώπιση πολλαπλών εμφανίσεων του ίδιου στοιχείου.

Στην πρώτη περίπτωση, η καθοριζόμενη XPath προσδιορίζει ένα μόνο στοιχείο στην ιστοσελίδα (δηλαδή, το κελί ενός πίνακα). Στην δεύτερη περίπτωση, η XPath προσδιορίζει πολλαπλές εμφανίσεις του ίδιου τύπου στοιχείου (ακόμα ένα κελί πίνακα) που μοιράζονται την ίδια ιεραρχική θέση. Για τους σκοπούς του Web Data Extraction, η δυνατότητα εκμετάλλευσης ενός τέτοιου ισχυρό εργαλείο έχει μεγάλη σημασία: η υιοθέτηση της XPath ως εργαλείο για την αντιμετώπιση στοιχείων στην ιστοσελίδα έχει αξιοποιηθεί σε μεγάλο βαθμό στη βιβλιογραφία.

Η μεγάλη αδυναμία της XPath σχετίζεται με την έλλειψη ελαστικότητας: κάθε έκφραση XPath είναι απολύτως σχετική με τη δομή της ιστοσελίδας πάνω στην οποία έχει οριστεί. Ωστόσο, οι συνέπειες αυτού του περιορισμού έχουν μετριαστεί μόνο εν μέρει, με την εισαγωγή σχετικών εκφράσεων μονοπατιού, στιες τελευταίες εκδόσεις (βλ. XPath 2.0: www.w3.org/TR/xpath20). Σε γενικές γραμμές, ακόμα και μικρές αλλαγές στη δομή μίας ιστοσελίδας μπορούν να διαβρώσουν την σωστή λειτουργία μίας έκφρασης XPath που έχει οριστεί σε μία προηγούμενη έκδοση της ιστοσελίδας. Για παράδειγμα, οι ιστοσελίδες που δημιουργούνται από ένα σενάριο εντολών. Υποθέτουμε ότι το σενάριο υφίσταται κάποια αλλαγή: μπορούμε να αναμένουμε ότι η δομή του δέντρου της σελίδας HTML που παράγεται από το εν λόγω σενάριο θα αλλάξει αναλόγως.

Για να διατηρηθεί η διαδικασία του Web Data Extraction λειτουργική, θα πρέπει να επικαιροποιηθεί η έκφραση κάθε φορά που γίνεται κάποια αλλαγή στο μοντέλο παραγωγής της υποκείμενης σελίδας. Μία τέτοια πράξη θα απαιτήσει υψηλό βαθμό ανθρώπινης εμπλοκής και, ως εκ τούτου, το κόστος της θα μπορούσε να είναι απαγορευτικά μεγάλο. Για το σκοπό αυτό, ορισμένοι συγγραφείς εισήγαγαν την έννοια του ισχυρού wrapper: πρότειναν μία στρατηγική για την εύρεση, ανάμεσα σε όλες τις εκφράσεις XPath που είναι ικανές για την εξαγωγή των ίδιων πληροφοριών από μια ιστοσελίδα, αυτήν που επηρεάζεται λιγότερο από τις εν δυνάμει αλλαγές στη δομή της ιστοσελίδας και μια τέτοια έκφραση προσδιορίζει το πιο ισχυρό wrapper (Dalvi, Bohannon, and Sha, 2009; Dalvi, Kumar, and Soliman, 2011).

Σε γενικές γραμμές, για να γίνει η όλη διαδικασία του Web Data Extraction ισχυρή, χρειαζόμαστε τα κατάλληλα εργαλεία που θα μας επιτρέψουν να μετρήσουμε τον βαθμό ομοιότητας των δύο εγγράφων. Ένα τέτοιο έργο μπορεί να πραγματοποιηθεί με την ανίχνευση δομικών παραλλαγών στα δέντρα DOM που σχετίζονται με τα έγγραφα. Μερικές τεχνικές που ονομάζονται στρατηγικές αντιστοίχησης δέντρων, αποτελούν καλή υποψηφιότητα για τον εντοπισμό ομοιοτήτων μεταξύ δύο δέντρων.

2.8 Αλγόριθμοι αντιστοίχησης της απόστασης διόρθωσης δέντρου (Tree edit distance)

Η πρώτη τεχνική που περιγράφουμε ονομάζεται αντιστοίχηση της

απόστασης διόρθωσης δέντρου (tree edit distance). Το πρόβλημα του υπολογισμού της απόστασης διόρθωσης δέντρου ανάμεσα στα δέντρα αποτελεί μία παραλλαγή του κλασικού προβλήματος του string edit distance. Για δύο επισημασμένα διατεταγμένα δέντρα με ρίζες A και B, το πρόβλημα είναι να βρεθεί μία αντιστοίχηση για να μετατραπεί το A σε B (ή το αντίστροφο) με τον ελάχιστο αριθμό ενεργειών. Το σύνολο των πιθανών ενεργειών αποτελείται από την διαγραφή κόμβου, την προσθήκη ή την αντικατάσταση. Σε κάθε ενέργεια μπορεί να εφαρμοστεί ένα κόστος, και στην περίπτωση αυτή, το έργο μετατρέπεται σε ένα πρόβλημα ελαχιστοποίησης κόστους (δηλαδή, η εύρεση της αλληλουχίας των ενεργειών με το ελάχιστο κόστος για το μετασχηματισμό του A σε B).

Κεφάλαιο 3ο Web wrappers

Στην βιβλιογραφία, οποιαδήποτε διαδικασία που στοχεύει στην εξαγωγή δεδομένων δομή από μη δομημένες (ή ημι-δομημένες) πηγές δεδομένων συνήθως αναφέρεται ως wrapper. Στο πλαίσιο του Web Data Extraction παρέχουμε τον ακόλουθο ορισμό: Μια διαδικασία, που θα μπορούσε να εφαρμόσει μία ή πολλές διαφορετικές κατηγορίες αλγορίθμων και η οποία αναζητά και βρίσκει τα δεδομένα που απαιτούνται από έναν ανθρώπινο χρήστη, αποσπώντας τα από μη δομημένες (ή ημι-δομημένες) πηγές στον Ιστό και τα μετατρέπει σε δομημένα δεδομένα, συγχωνεύοντας και ενοποιώντας τις πληροφορίες αυτές για περαιτέρω επεξεργασία, με έναν ημι-αυτόματο ή πλήρως αυτόματο τρόπο. Οι wrappers Ιστού χαρακτηρίζονται από έναν κύκλο ζωής που αποτελείται από διάφορα στάδια:

1. Παραγωγή Wrapper: ο wrapper ορίζεται σύμφωνα με κάποια τεχνική(ες)
2. Εκτέλεση Wrapper: ο wrapper τρέχει και εξάγει πληροφορίες συνεχώς
3. Συντήρηση Wrapper: η δομή των πηγών δεδομένων μπορεί να αλλάξει και ο wrapper πρέπει να προσαρμοστεί ανάλογα για να συνεχίσει να δουλεύει κανονικά.

Τα δύο πρώτα βήματα του κύκλου ζωής ενός wrapper, η παραγωγή και η εκτέλεσή του, μπορούν να εφαρμοστούν χειρωνακτικά. Για παράδειγμα με τον καθορισμό και την εκτέλεση τακτικών εκφράσεων πάνω από τα έγγραφα HTML. Εναλλακτικά, κάτι που είναι και ο στόχος των συστημάτων Web Data Extraction, οι wrappers μπορούν να οριστούν και να εκτελούνται χρησιμοποιώντας μια επαγωγική προσέγγιση (wrapper induction) (Kushmerick, 1997). Η επαγωγή wrapper ιστού είναι δύσκολη επειδή απαιτεί στρατηγικές αυτοματισμού υψηλού επιπέδου. Υπάρχουν επίσης υβριδικές προσεγγίσεις που καθιστούν δυνατό για τους χρήστες να δημιουργήσουν και να τρέξουν wrappers ημι-αυτόματα μέσω οπτικών διεπαφών.

Το τελευταίο βήμα του κύκλου ζωής ενός wrapper είναι η συντήρηση: Οι ιστοσελίδες αλλάζουν τη δομή τους συνεχώς και χωρίς προειδοποίηση. Αυτό ενδέχεται να καταστρέψει την σωστή λειτουργία ενός wrapper ιστού, του οποίου ο ορισμός είναι συνήθως στενά συνδεδεμένος με τη δομή των ιστοσελίδων που έχει εγκριθεί για την

παραγωγή του. Ο καθορισμός αυτόματων στρατηγικών για τη συντήρηση του wrapper είναι εξέχουσας σημασίας για τη διασφάλιση της ορθότητας των δεδομένων που εξήχθησαν και την ευρωστία των πλατφορμών Web Data Extraction.

3.1 Δημιουργία και εκτέλεση wrapper

Το πρώτο βήμα στον κύκλο ζωής των wrapper είναι η δημιουργία τους. Οι αρχικές πλατφόρμες Web Data Extraction παρείχαν μόνο υποστήριξη για την χειρωνακτική δημιουργία των wrappers, η οποία απαιτούσε ανθρώπινη εμπειρία και δεξιότητες σε γλώσσες προγραμματισμού για την σύνταξη αρχείου εντολών σε θέση για να προσδιορίσει και να εξάγει επιλεγμένα κομμάτια πληροφοριών σε μια ιστοσελίδα.

Στα τέλη της δεκαετίας του ’90 εμφανίστηκαν πιο προηγμένα συστήματα εξαγωγής δεδομένων στον Ιστό. Το κύριο χαρακτηριστικό τους ήταν η δυνατότητα που έδιναν στους χρήστες να καθορίζουν και να εκτελούν Web wrappers μέσω διαδραστικών γραφικών διεπαφών χρηστών (GUIs). Στις περισσότερες περιπτώσεις, δεν απαιτούνταν καμία βαθιά κατανόηση της γλώσσας προγραμματισμού wrappers, καθώς τα wrappers δημιουργούνταν αυτόματα (ή ημιαυτόματα) από το σύστημα αξιοποίησης οδηγιών που παρέχονταν από τους χρήστες μέσω της πλατφόρμας διεπαφής.

Στη συνέχεια εξετάζονται λεπτομερώς τρεις τύποι προσέγγισης που διέπουν αυτού του είδους τις πλατφόρμες: regular expressions (κανονικές εκφράσεις), wrapper programming languages (γλώσσες προγραμματισμού wrapper) και tree-based (προσέγγιση με βάση το δέντρο).

Προσέγγιση που βασίζεται στις κανονικές εκφράσεις. Μία από τις πιο κοινές προσεγγίσεις βασίζεται στις κανονικές εκφράσεις, οι οποίες αποτελούν μία ισχυρή επίσημη γλώσσα που χρησιμοποιείται για τον εντοπισμό σειρών ή μοτίβων σε αδόμητα κείμενα βάσει κάποιων κριτηρίων που να ταιριάζουν. Οι κανόνες θα μπορούσαν να είναι πολύπλοκοι επίσης, οπότε, το να τους γράψει κανείς με το χέρι απαιτεί πολύ χρόνο και μεγάλη εμπειρία: τα wrappers που βασίζονται σε κανονικές εκφράσεις δημιουργούν δυναμικά κανόνες για την εξαγωγή των επιθυμητών δεδομένων από τις ιστοσελίδες. Συνήθως, η σύνταξη κανονικών εκφράσεων στις σελίδες HTML βασίζεται στα ακόλουθα κριτήρια: όρια των λέξεων, ετικέτες HTML, δομή πινάκων, κ.λπ.

Ένα αξιοσημείωτο εργαλείο εφαρμογής για την εξαγωγή βάσει κανονικών εκφράσεων είναι το W4F (Sahuguet & Azavant, 1999). Το W4F υιοθετεί μια προσέγγιση επισημείωσης: αντί να θέσει σε δοκιμασία τους χρήστες να ασχοληθούν με την σύνταξη εγγράφων HTML, το W4F διευκολύνει το σχεδιασμό του wrapper μέσω μιας διαδικασίας οδηγού. Αυτός ο οδηγός επιτρέπει στους χρήστες να επιλέγουν και να επισημειώνουν τα στοιχεία απευθείας στην ιστοσελίδα. Το W4F παράγει τους κανόνες εξαγωγής των κανονικών εκφράσεων των επισημειωμένων στοιχείων και τα παρέχει στους χρήστες. Ένα περαιτέρω βήμα, που είναι η βελτιστοποίηση των κανονικών εκφράσεων που παράγονται από το W4F,

ανατίθεται σε πιο έμπειρους χρήστες - στην πραγματικότητα, το εργαλείο δεν είναι πάντα σε θέση να παρέχει τον καλύτερο κανόνα εξαγωγής.

Με την πλήρη εκμετάλλευση της δύναμης των κανονικών εκφράσεων, οι κανόνες εξαγωγής του W4F περιλαμβάνουν τις εκφράσεις match και split, η οποία διαχωρίζει τις λέξεις, με την επισημείωση διάφορων στοιχείων στην ίδια σειρά. Το μειονέκτημα της υιοθέτησης των κανονικών εκφράσεων είναι η έλλειψη ευελιξίας. Για παράδειγμα, όταν προκύπτει ακόμη και μια μικρή αλλαγή στη δομή ή το περιεχόμενο μιας ιστοσελίδας, κάθε κανονική έκφραση είναι πολύ πιθανό να σταματήσει να λειτουργεί και θα πρέπει να ξαναγραφτεί. Η διαδικασία αυτή συνεπάγεται μια μεγάλη δέσμευση από ανθρώπινους χρήστες, ιδίως για τη συντήρηση των συστημάτων που βασίζονται στις κανονικές εκφράσεις. Για τους λόγους αυτούς έχουν αναπτυχθεί πιο ευέλικτες και ισχυρές γλώσσες για να αυξήσουν τις δυνατότητες των πλατφορμών Web Data Extraction.

Λογική προσέγγιση. Ένα παράδειγμα ισχυρών γλωσσών που αναπτύχθηκε για τους σκοπούς της εξαγωγής δεδομένων προέρχεται από το τις γλώσσες προγραμματισμού wrappers ειδικά τον Ιστό. Τα εργαλεία που βασίζονται σε γλώσσες προγραμματισμού wrapper δεν θεωρούν τις ιστοσελίδες ως απλές συμβολοσειρές κειμένου, αλλά ως ημιδομημένα έγγραφα δέντρου, ενώ η DOM της ιστοσελίδας αποτελεί την δομή του, όπου οι κόμβοι είναι στοιχεία που χαρακτηρίζονται τόσο από τις ιδιότητές τους, όσο και από το περιεχόμενό τους. Το πλεονέκτημα μιας τέτοιας προσέγγισης είναι ότι οι γλώσσες προγραμματισμού των wrappers μπορούν να οριστούν ώστε να εκμεταλλευτούν πλήρως τόσο την ημιδομημένη φύση των ιστοσελίδων, όσο και του περιεχομένου τους - η πρώτη πτυχή στερείται σε συστήματα που βασίζονται σε κανονικές εκφράσεις.

Η πρώτη ισχυρή γλώσσα wrapper έχει επισημοποιηθεί από τους Gottlob και Koch (2004). Οι λειτουργίες εξαγωγής πληροφοριών που υλοποιούνται από αυτήν την γλώσσα για wrappers βασίζονται σε μοναδιαίους datalogs πάνω από τα δέντρα. Οι συγγραφείς απέδειξαν ότι οι μοναδιαίοι datalogs πάνω από το δέντρο ισοδυναμούν με την μοναδιαία λογική δεύτερης τάξης (MSO) και ως εκ τούτου είναι πολύ εκφραστικά. Ωστόσο, σε αντίθεση με την MSO, ένα wrapper σε μοναδιαίους datalogs μπορεί να μοντελοποιηθεί όμορφα με έναν οπτικό και διαδραστικό τρόπο. Αυτό κάνει αυτή την γλώσσα του wrapper κατάλληλη για να ενσωματωθεί σε οπτικά εργαλεία, ικανοποιώντας την προϋπόθεση ότι όλα τα κατασκευάσματά του θα μπορούν να εφαρμοστούν μέσω των αντίστοιχων οπτικών αρχετύπων.

Ξεκινώντας από το αβαθμίδωτο ετικετοποιημένο δέντρο που αντιπροσωπεύει το DOM της ιστοσελίδας, ο αλγόριθμος ετικετοποιεί εκ νέου τους κόμβους, περικόπτει τους άσχετους και τελικά επιστρέφει ένα υποσύνολο κόμβων αρχικών δέντρων, που αντιπροσωπεύουν τα επιλεγμένα δεδομένα που εξάγονται. Η πρώτη εφαρμογή αυτής της γλώσσας για wrapper σε πραγματικά σενάρια οφείλεται στους Baumgartner et al. (2001). Ανέπτυξαν την γλώσσα Elog για wrapping που υλοποιεί σχεδόν όλες τις λειτουργίες εξαγωγής πληροφοριών της μοναδιαίας datalog με κάποιους μικρούς περιορισμούς.

Η γλώσσα Elog χρησιμοποιείται ως κύρια μέθοδος εξαγωγής του συστήματος Lixto Visual Wrapper. Αυτή η πλατφόρμα παρέχει ένα γραφικό περιβάλλον χρήστη για να επιλέξει μέσα από οπτικές προδιαγραφές, τα πρότυπα στις ιστοσελίδες σε ιεραρχική σειρά, τονίζοντας τα στοιχεία του εγγράφου και καθορίζοντας τις σχέσεις μεταξύ τους. Οι πληροφορίες που προσδιορίζονται με τον τρόπο αυτό θα μπορούσαν να είναι υπερβολικά γενικές, έτσι το σύστημα επιτρέπει στους χρήστες να προσθέτουν κάποια όρια περιορισμών, για παράδειγμα, before/after, not-before/not-after, εσωτερικά όρια και όρια για το εύρος. Τέλος, τα επιλεγμένα δεδομένα μεταφράζονται σε έγγραφα XML, χρησιμοποιώντας μοτίβα ονομάτων όπως τα ονόματα στοιχείων XML, λαμβάνοντας δομημένα δεδομένων από ημι-δομημένες ιστοσελίδες.

Προσέγγιση με βάση το δέντρο [μερική ευθυγράμμιση δέντρου]. Η τελευταία τεχνική σχετίζεται με την δημιουργία wrapper και ονομάζεται μερική ευθυγράμμιση δέντρου. Επισημοποιήθηκε πρόσφατα από τους Zhai και Liu (2005) οι οποίοι επίσης ανέπτυξαν ένα σύστημα Web Data Extraction με βάση αυτήν την τεχνική.

Η τεχνική αυτή βασίζεται στην ιδέα ότι οι πληροφορίες στα έγγραφα του Ιστού συνήθως συλλέγονται σε όμορες περιοχές μιας σελίδας, που ονομάζονται περιοχές εγγραφής δεδομένων. Η στρατηγική της μερικής ευθυγράμμισης δέντρου συνίσταται στον εντοπισμό και την εξαγωγή αυτών των περιοχών. Πιο συγκεκριμένα, οι συγγραφείς εμπνεύστηκαν από τους αλγορίθμους αντιστοίχησης δέντρων, χρησιμοποιώντας την αντιστοίχηση της απόστασης διόρθωσης δέντρου (tree edit distance). Ο αλγόριθμος λειτουργεί σε δύο στάδια:

1. Τμηματοποίηση

2. Μερική ευθυγράμμιση δέντρου

Στην πρώτη φάση, η ιστοσελίδα χωρίζεται σε τμήματα, χωρίς την εξαγωγή δεδομένων. Αυτή η φάση προ-επεξεργασίας είναι καθοριστικής σημασίας για το τελευταίο βήμα. Στην πραγματικότητα, το σύστημα δεν εκτελεί μόνο μια ανάλυση του εγγράφου της ιστοσελίδας βάσει του δέντρου DOM, αλλά στηρίζεται και σε οπτικά ερεθίσματα (όπως και στη χωρική τεχνική συλλογιστική), προσπαθώντας να εντοπίσει τα κενά μεταξύ των αρχείων δεδομένων. Αυτό το βήμα είναι χρήσιμο καθώς βοηθά και στην διαδικασία της εξαγωγής δομημένων πληροφοριών από το έγγραφο HTML, σε περιπτώσεις που γίνεται κατάχρηση της σύνταξης HTML, για παράδειγμα με την χρήση δομής πίνακα αντί του CSS για την οργάνωση της γραφικής πλευράς της σελίδας.

Στο δεύτερο στάδιο, ο αλγόριθμος μερικής ευθυγράμμισης δέντρου εφαρμόζεται σε αρχεία δεδομένων που έχουν εντοπιστεί νωρίτερα. Κάθε εγγραφή δεδομένων εξάγεται από την θέση της στο υποδέντρο DOM, αποτελώντας την ρίζα ενός νέου, ενιαίου δέντρου. Αυτό επειδή κάθε εγγραφή δεδομένων θα μπορούσε να περιέχεται σε περισσότερα από ένα μη-συνεχόμενα υπο-δέντρα στο αρχικό δέντρο DOM. Η προσέγγιση της μερικής ευθυγράμμισης δέντρου συνεπάγεται την ευθυγράμμιση των πεδίων δεδομένων με βεβαιότητα, αποκλείοντας εκείνα που δεν μπορούν να ευθυγραμμιστούν ώστε να εξασφαλιστεί ένας υψηλός βαθμό ακριβείας.

Κατά τη διάρκεια αυτής της διαδικασίας δεν εμπλέκονται καθόλου στοιχεία δεδομένων, διότι η μερική ευθυγράμμιση δέντρου λειτουργεί μόνο στην αντιστοίχιση ετικετών δέντρου, οι οποίες εκπροσωπούν το ελάχιστο δυνατό κόστος, από την άποψη των εργασιών (δηλαδή, την αφαίρεση κόμβου, την εισαγωγή κόμβου, την αντικατάσταση κόμβου) για να μετατραπεί ένας κόμβος σε έναν άλλο. Το μειονέκτημα αυτού του χαρακτηριστικού του αλγορίθμου είναι ότι οι επιδόσεις για την ανάκλησή του (δηλαδή, η δυνατότητα ανάκτησης όλων των προσδοκώμενων πληροφοριών) μπορεί να φθίνει σε περίπτωση πολύπλοκων δομών εγγράφου HTML. Επιπλέον, και στην περίπτωση της μερικής ευθυγράμμισης δέντρου, η λειτουργία αυτής της στρατηγικής είναι αυστηρά σχετική με τη δομή της ιστοσελίδας κατά τη στιγμή του ορισμού της ευθυγράμμισης. Αυτό σημαίνει ότι η μέθοδος είναι πολύ ευαίσθητη ακόμη και σε μικρές αλλαγές, που θα μπορούσαν να θέσουν σε κίνδυνο την λειτουργία του αλγορίθμου και την ορθή εξαγωγή των πληροφοριών. Ακόμη και σε αυτή την προσέγγιση, προκύπτει το πρόβλημα της συντήρησης με εξαιρετική σημασία.

Προσεγγίσεις Μηχανικής Μάθησης. Οι τεχνικές μηχανικής μάθησης ταιριάζουν αρκετά με τον σκοπό της εξαγωγής συγκεκριμένων με τον τομέα πληροφοριών από πηγές του Ιστού, καθώς βασίζονται σε συνεδρίες εκμάθησης κατά τις οποίες ένα σύστημα αποκτά εμπειρία στον τομέα.

Οι προσεγγίσεις της μηχανικής μάθησης απαιτούν μία βαθμίδα εκμάθησης όπου οι ειδικοί του τομέα παρέχουν με το χέρι κάποιες ιστοσελίδες με ετικέτες, οι οποίες αποκτήθηκαν από διάφορους ιστοτόπους, αλλά και από την ίδια την ιστοσελίδα. Ιδιαίτερη προσοχή πρέπει να δοθεί στην παροχή παραδειγμάτων ιστοσελίδων που ανήκουν στον ίδιο τομέα, αλλά που παρουσιάζουν διαφορετικές δομές.

Αυτό, γιατί, ακόμη και στο ίδιο σενάριο τομέα, τα πρότυπα που συνήθως θεσπίζονται για τη δημιουργία ιστοσελίδων με δυναμικό περιεχόμενο, διαφέρουν και το σύστημα θα πρέπει να είναι ικανό να μάθει πώς να εξαγάγει τις πληροφορίες σε αυτά τα πλαίσια. Τα συστήματα Στατιστικής Μηχανικής Μάθησης αναπτύχθηκαν επίσης στηριζόμενα σε μοντέλα συνθηκών ή στην προσαρμοστική αναζήτηση ως την εναλλακτική λύση στην ανθρώπινη γνώση και την αλληλεπίδραση. Στη συνέχεια περιγράφονται εν συντομία μερικές προσεγγίσεις Web Data Extraction που βασίζονται σε αλγορίθμους μηχανικής μάθησης (Muslea et al., 1999).

Μία από τις πρώτες προσεγγίσεις είναι το WIEN (Kushmerick, 2000). Το WIEN βασίστηκε σε διάφορες τεχνικές επαγωγικής μάθησης και ήταν σε θέση να ετικετοποιεί αυτόματα σελίδες εκπαίδευσης, που αντιπροσωπεύουν εκ των πραγμάτων ένα υβριδικό σύστημα του οποίου η διαδικασία εκπαίδευσης συνεπάγεται χαμηλή ανθρώπινη εμπλοκή. Η άλλη πλευρά της υψηλής αυτοματοποίησης του WIEN ήταν ο μεγάλος αριθμός των περιορισμών που συνδέονται με το επαγωγικό του σύστημα: για παράδειγμα, η διαδικασία εξαγωγής δεδομένων δεν ήταν σε θέση να αντιμετωπίσει τις τιμές που έλειπαν, μία περίπτωση που συμβαίνει συχνά και θέτει σοβαρούς περιορισμούς στην προσαρμοστικότητα του WIEN σε πραγματικά σενάρια.

Το Rapier (Robust Automated Production of Information Extraction

Rules) (Mooney, 1999) είναι ένα σύστημα που έχει σχεδιαστεί για την εκμάθηση κανόνων για την εξαγωγή πληροφοριών από έγγραφα και το κύριο πλεονέκτημά του είναι, ίσως, η ικανότητα εκμάθησης των κανόνων αυτών απευθείας από τα έγγραφα χωρίς προηγούμενη ανάλυση ή οποιαδήποτε προεπεξεργασία. Οι κανόνες εξαγωγής είναι σχετικά απλοί και κάνουν χρήση των περιορισμένων συντακτικών και σημασιολογικών πληροφοριών.

Από τη μία πλευρά, οι κανόνες του Rapier είναι ευέλικτοι, επειδή δεν περιορίζονται στο να περιέχουν ένα σταθερό αριθμό λέξεων, αλλά από την άλλη πλευρά, είναι δύσκολο να αναγνωρίσει κανείς ποιοι κανόνες είναι πραγματικά χρήσιμοι για την εκτέλεση της εξαγωγής δεδομένων. Για το σκοπό αυτό, ένας αλγόριθμος εκμάθησης έχει αναπτυχθεί για να εντοπιστούν οι αποτελεσματικοί κανόνες και ο αλγόριθμος αυτός βασίζεται στον Επαγωγικό Λογικό Προγραμματισμό.

Το WHISK (Soderland, 1999) στηρίζεται σε ένα αλγόριθμο εποπτευόμενης εκμάθησης που δημιουργεί κανόνες για την εξαγωγή πληροφοριών από έγγραφα κειμένου. Το WHISK είναι σε θέση να χειριστεί ένα ευρύ φάσμα εγγράφων κειμένου που κυμαίνονται από εξαιρετικά δομημένα έγγραφα (όπως HTML) μέχρι και ελεύθερα κείμενα. Οι κανόνες εξαγωγής στο WHISK μπορούν να θεωρηθούν ως ένα ιδιαίτερο είδος κανονικών εκφράσεων με τα εξής δύο στοιχεία: το πρώτο καθορίζει το πλαίσιο στο οποίο μια έκφραση πρέπει να θεωρηθεί συναφής και το δεύτερο καθορίζει τα ακριβή διαχωριστικά όρια της φράσης που πρόκειται να εξαχθεί (π.χ. τα όρια του κειμένου που πρέπει να εξαχθεί). Ανάλογα με τη δομή ενός εγγράφου, το WHISK δημιουργεί κανόνες που βασίζονται ακριβώς πάνω σε ένα από τα δύο στοιχεία που αναφέρθηκαν παραπάνω.

Ειδικότερα, στην περίπτωση του ελεύθερου κειμένου χρησιμοποιεί κανόνες που βασίζονται στα συμφραζόμενα, ενώ στην περίπτωση των δομημένων κειμένων χρησιμοποιεί διαχωριστικά όρια. Επιπλέον, για όλα αυτά τα έγγραφα των οποίων η δομή κυμαίνεται μεταξύ δομημένου εγγράφου και ελεύθερου κειμένου, το WHISK είναι σε θέση να χρησιμοποιήσει έναν συνδυασμό των κανόνων ως προς τα συμφραζόμενα και κανόνες με βάση τα διαχωριστικά όρια. Το WHISK χρησιμοποιεί έναν αλγόριθμο εποπτευόμενης μάθησης για την επαγωγή νέων κανόνων από ένα σύνολο περιπτώσεων που έχουν επισημανθεί με το χέρι. Προκειμένου να διατηρηθεί η ανθρώπινη εμπλοκή περιορισμένη, το WHISK παρεμβάλλει την εκμάθηση των νέων κανόνων και την επισημείωση νέων περιπτώσεων. Ως εκ τούτου, η διαδικασία της εκμάθησης/επισημείωσης είναι επαναληπτική και αντί να παρουσιάζει αυθαίρετες περιπτώσεις, το WHISK παρουσιάζει περιπτώσεις που είναι κοντά στα παραδείγματα που μπορούν να αντιμετωπιστούν με τους κανόνες τους οποίους το WHISK έχει μάθει μέχρι τώρα.

Το SRV προτάθηκε από τον Freytag (2000). Το SRV λαμβάνει ως εισαγωγή δεδομένων ένα σύνολο εγγράφων με ετικέτες και εξάγει ορισμένα χαρακτηριστικά που περιγράφουν τις λεκτικές μονάδες που μπορούν να εξαχθούν από ένα έγγραφο. Τα χαρακτηριστικά κατατάσσονται ως απλά, αν αντιστοιχούν μία λεκτική μονάδα σε μια κατηγορική τιμή και ως σχεσιακά αν αντιστοιχούν μία λεκτική μονάδα σε

μία άλλη λεκτική μονάδα. Το SRV είναι επίσης σε θέση να διαχειριστεί τα χαρακτηριστικά που κωδικοποιούν την δομική πτυχή του εγγράφου (π.χ., αν μία λεκτική μονάδα είναι ένα ρήμα). Οι κανόνες εξαγωγής μπορούν να εκφράζονται επί τη βάσει των διαθέσιμων χαρακτηριστικών. Για την δημιουργία νέων κανόνων, το SRV χρησιμοποιεί έναν ταξινομητή Naive Bayes, σε συνδυασμό με έναν σχεσιακό μαθητευόμενο.

Το SoftMealy (Hsu & Dung, 1998) ήταν το πρώτο σύστημα εισαγωγής wrapper, ειδικά σχεδιασμένο για να λειτουργεί στο πλαίσιο του Web Data Extraction. Στηριζόμενο σε καταστάσεις με μη-ντετερμινιστικά πεπερασμένα αυτόματα (FST), το SoftMealy χρησιμοποιεί μια προσέγγιση επαγωγικής μάθησης από κάτω προς τα πάνω για την εκμάθηση των κανόνων εξαγωγής. Κατά τη διάρκεια της εκμάθησης, το σύστημα αποκτά σελίδες εκμάθησης που παρουσιάζονται ως ένα αυτόματο με όλες τις δυνατές μεταθέσεις των ιστοσελίδων: οι καταστάσεις αντιπροσωπεύουν τα εξαγόμενα δεδομένα, ενώ οι καταστάσεις των μεταβάσεων αντιπροσωπεύουν τους κανόνες εξαγωγής. Η κύρια δύναμη του SoftMealy ήταν ότι εισήγαγε μία νέα μέθοδο για την εσωτερική αναπαράσταση των εγγράφων HTML. Πιο αναλυτικά, κατά το στάδιο της προεπεξεργασίας, κάθε εξεταζόμενη ιστοσελίδα κωδικοποιούνταν σε λεκτικές μονάδες (που ορίζονταν σύμφωνα με ένα σύνολο επαγωγικών κανόνων). Στη συνέχεια, οι λεκτικές μονάδες αξιοποιούνταν για τον καθορισμό διαχωριστών, οι οποίοι θεωρούνταν ως τα αόρατα όρια ανάμεσα σε δύο διαδοχικές λεκτικές μονάδες. Τέλος, το FST τροφοδοτούνταν από την αλληλουχία των διαχωριστών, αντί των πρώτων συμβολοσειρών HTML (όπως στο WIEN), έτσι ώστε οι λεκτικές μονάδες να ταιριάζουν με τους κανόνες για τα συμφραζόμενα (που έχουν οριστεί για να χαρακτηρίσουν ένα σύνολο μεμονωμένων διαχωριστών) για να καθοριστούν οι καταστάσεις μετάβασης. Τα πλεονεκτήματα του SoftMealy σε σχέση με το WIEN είναι αξιοσημείωτα: στην πραγματικότητα, το σύστημα ήταν σε θέση να ασχολείται με μια σειρά από εξαιρέσεις, όπως χαμένες τιμές/γνωρίσματα, γνωρίσματα πολλαπλών τιμών, διάφορες μεταθέσεις στα γνωρίσματα και επίσης με τα ορθογραφικά λάθη.

Το τελευταίο σύστημα που βασίζεται στην εκμάθησης ονομάζεται STALKER (Muslea et al., 1999). Ήταν ένα εποπτευόμενο σύστημα εκμάθησης για την επαγωγή wrapper, το οποίο μοιράζεται κάποιες ομοιότητες με το SoftMealy. Η κύρια διαφορά μεταξύ αυτών των δύο συστημάτων είναι η προδιαγραφή των σχετικών δεδομένων: στο STALKER, ένα σύνολο από λεκτικές μονάδες τοποθετούνται στην ιστοσελίδα με το χέρι, έτσι ώστε να εντοπιστούν στοιχεία που ο χρήστης προτίθεται να εξάγει. Η πτυχή αυτή διασφαλίζει την ικανότητα του STALKER στην αντιμετώπιση της ύπαρξης κενών τιμών, ιεραρχικές δομές και με αντικείμενα για τα οποία δεν έχει δοθεί εντολή. Αυτό το σύστημα μοντελοποιεί το περιεχόμενο μιας ιστοσελίδας μέσω ιεραρχικών σχέσεων, το οποίο παρουσιάζεται με την χρήση μιας δομής δεδομένων δέντρου που ονομάζεται embedded catalog tree (EC tree). Η ρίζα του δέντρου EC περιλαμβάνει την ακολουθία όλων των λεκτικών μονάδων (ενώ το STALKER θεωρεί ως λεκτική μονάδα κάθε κομμάτι του κειμένου ή HTML tag στο έγγραφο).

Κάθε κόμβος-παιδί είναι μία υπο-ακολουθία των λεκτικών μονάδων

που κληρονόμησε από τον γονικό κόμβο. Αυτό σημαίνει ότι κάθε γονικός κόμβος είναι μία υπερ-ακολουθία λεκτικών μονάδων των παιδιών του. Η υπερ-ακολουθία χρησιμοποιείται σε κάθε επίπεδο της ιεραρχίας, για την παρακολούθηση του περιεχομένου στα υπο-επίπεδα του δένδρου EC. Η εξαγωγή των στοιχείων που παρουσιάζουν ενδιαφέρον για το χρήστη επιτυγχάνεται συνάγοντας μία σειρά κανόνων εξαγωγής στο ίδιο το δέντρο EC, ένα τυπικό παράδειγμα του κανόνα εξαγωγής που συνάγεται από το STALKER είναι η κατασκευή SkipTo(T), μία οδηγία που υποδεικνύει, κατά τη φάση της εξαγωγής, την παράλειψη όλων των λεκτικών μονάδων μέχρι να βρεθεί η πρώτη εμφάνιση της λεκτικής μονάδας T. Η επαγωγή των κανόνων εξαγωγής εκμεταλλεύεται την έννοια των οροσήμων, των ακολουθιών των διαδοχικών λεκτικών μονάδων που εγκρίθηκαν για τον εντοπισμό της αρχής και του τέλους ενός συγκεκριμένου στοιχείου για εξαγωγή. Το STALKER είναι επίσης σε θέση να καθορίσει wildcards, τάξεις των γενικών λεκτικών μονάδων που δεν αποκλείουν τις περισσότερο συγκεκριμένες λεκτικές μονάδες.

3.2 Το πρόβλημα της συντήρησης του wrapper

Η δημιουργία wrapper, ανεξάρτητα από την χρησιμοποιούμενη τεχνική, αποτελεί μία πτυχή του προβλήματος της εξαγωγής δεδομένων από πηγές στον Ιστό. Από την άλλη πλευρά, η συντήρηση του wrapper είναι εξίσου σημαντική, έτσι ώστε η πλατφόρμα εξαγωγής των δεδομένων ιστού να μπορεί να φτάσει σε υψηλά επίπεδα σταθερότητας και αξιοπιστίας, μαζί με το επίπεδο του αυτοματισμού και το χαμηλό επίπεδο της ανθρώπινης εμπλοκής. Στην πραγματικότητα, σε αντίθεση με τα στατικά έγγραφα, οι ιστοσελίδες αλλάζουν δυναμικά και εξελίσσονται και η δομή τους μπορεί να αλλάξει, μερικές φορές με αποτέλεσμα τα wrappers που έχουν καθοριστεί από πριν, να μην είναι πλέον σε θέση να εξάγουν με επιτυχία τα δεδομένα.

Υπό το φως αυτών των υποθέσεων, θα μπορούσε κανείς να θέσει το επιχείρημα ότι η διατήρηση του wrapper αποτελεί ένα κρίσιμο βήμα στην διαδικασία της εξαγωγής δεδομένων ιστού. Αν και αυτή η πτυχή δεν έχει λάβει αρκετή προσοχή στην βιβλιογραφία (πολύ λιγότερο από ότι το πρόβλημα της δημιουργίας wrapper) τουλάχιστον μέχρι τώρα τελευταία.

Κατά το αρχικό στάδιο, στην πραγματικότητα η συντήρηση του wrapper πραγματοποιούνταν με το χέρι: οι χρήστες που συνήθως σχεδιάζουν Web wrappers επικαιροποιούσαν ή ξαναέγραφαν αυτά τα wrappers κάθε φορά που η δομή μιας συγκεκριμένης ιστοσελίδας τροποποιούνταν. Η προσέγγιση του εγχειριδίου συντήρησης ταιριάζει αρκετά καλά για την αντιμετώπιση μικρών προβλημάτων, αλλά καθίσταται ακατάλληλη όταν αυξάνει σε μεγάλο βαθμό ο αριθμός των ιστοσελίδων. Δεδομένου ότι στα σενάρια των επιχειρήσεων οι συνηθισμένες εργασίες εξαγωγής δεδομένων μπορεί να περιλαμβάνουν χιλιάδες (ή ακόμη περισσότερες) ιστοσελίδες, που δημιουργούνται δυναμικά και επικαιροποιούνται συχνά, το εγχειρίδιο συντήρησης wrapper δεν αποτελεί πλέον εφικτή λύση στις πραγματικές εφαρμογές.

Για τους λόγους αυτούς, το πρόβλημα της αυτοματοποίησης της συντήρησης του wrapper έχει αντιμετωπίσει από την πρόσφατη

βιβλιογραφία. Για παράδειγμα, η πρώτη προσπάθεια στην κατεύθυνση της αυτόματης συντήρησης wrapper έχει υποβληθεί από τον Kushmerick (2002), ο οποίος όρισε πρώτος την έννοια της επαλήθευσης του wrapper. Το θέμα της επαλήθευσης του wrapper προκύπτει ως απαιτούμενο βήμα κατά την εκτέλεση του wrapper, στην οποία ένα σύστημα Εξαγωγής Δεδομένων Ιστού αξιολογεί εάν τα καθορισμένα Web wrappers λειτουργούν σωστά ή, εναλλακτικά, αν η λειτουργία τους είναι κατεστραμμένη λόγω τροποποιήσεων στην δομή των υποκείμενων σελίδων. Ακολούθως, ο συγγραφέας εξέτασε κάποιες τεχνικές ημιαυτόματης συντήρησης wrapper για τον χειρισμό απλών προβλημάτων.

Η πρώτη μέθοδος που επιχειρεί να αυτοματοποιήσει την διαδικασία συντήρησης wrapper έχει αναπτυχθεί από τους Meng et al (2003). και η οποία ονομάζεται schema-guided wrapper maintenance. Στηρίζεται στον καθορισμό των σχεδίων XML κατά τη φάση της δημιουργίας του wrapper για να αξιοποιηθεί για την συντήρηση κατά τη διάρκεια του σταδίου εκτέλεσης.

Πιο πρόσφατα, οι Ferrara και Baumgartner (2011) ανέπτυξαν ένα σύστημα αυτόματης προσαρμογής wrapper (ένα είδος συντήρησης που πραγματοποιείται για να τροποποιήσει τα Web wrappers σύμφωνα με τη νέα δομή των ιστοσελίδων) το οποίο στηρίζεται στην ανάλυση των δομικών ομοιοτήτων μεταξύ των διαφόρων εκδοχών της ίδια ιστοσελίδας, χρησιμοποιώντας έναν αλγόριθμο επεξεργασίας δέντρου.

Schema-guided wrapper maintenance. Η πρώτη προσπάθεια για την αντιμετώπιση του προβλήματος της συντήρησης του wrapper με την παροχή υψηλού επιπέδου αυτοματοποίησης προήλθε από τους Meng et al.(2003). Οι συγγραφείς ανέπτυξαν την SG-WRAM (Schema-Guided Wrapper Maintenance), μια στρατηγική για την εξαγωγή δεδομένων ιστού με βάση την υπόθεση που στηρίζεται σε εμπειρικές παρατηρήσεις, ότι οι αλλαγές στις ιστοσελίδες, ακόμη και αν είναι μεγάλες, συχνά διατηρούν:

- Συντακτικά χαρακτηριστικά: τα συντακτικά χαρακτηριστικά των στοιχείων δεδομένων, όπως τα πρότυπα δεδομένων, τα μήκη των συμβολοσειρών, κλπ, τα οποία διατηρούνται ως επί το πλείστον.
- Υπερσυνδέσεις (Hyperlinks): Τα έγγραφα HTML συχνά εμπλουτίζονται με υπερσυνδέσεις που σπάνια απομακρύνονται με τις μετέπειτα τροποποιήσεις των ιστοσελίδων.
- Επισημειώσεις (Annotations): συνήθως διατηρούνται οι περιγραφικές πληροφορίες που αντιπροσωπεύουν την σημασιολογική έννοια ενός τμήματος πληροφοριών στο πλαίσιό του.

Με βάση αυτές τις υποθέσεις, οι συγγραφείς ανέπτυξαν ένα σύστημα εξαγωγής δεδομένων ιστού που, κατά την φάση της δημιουργίας wrapper, δημιουργεί συστήματα τα οποία χρησιμοποιούνται στην φάση της συντήρησης του wrapper. Πιο αναλυτικά, κατά τη διάρκεια της δημιουργίας του wrapper, ο χρήστης παρέχει έγγραφα HTML και σχήματα XML, καθορίζοντας μια αντιστοίχιση μεταξύ τους. Αργότερα, το σύστημα θα δημιουργήσει κανόνες εξαγωγής και, στην συνέχεια, θα τρέξει τα wrappers για την εξαγωγή δεδομένων, την δημιουργώντας ένα έγγραφο XML σύμφωνα με το καθορισμένο σχήμα XML. Κατά την φάση εκτέλεσης

του wrapper, εισάγεται ένα πρόσθετο συστατικό στην διαδικασία εξαγωγής δεδομένων: ο συντηρητής του wrapper.

Ο συντηρητής wrapper πραγματοποιεί ελέγχους για κάθε πιθανό ζήτημα στην εξαγωγή δεδομένων και δημιουργεί ένα αυτόματο πρωτόκολλο επισκευής για τα wrappers που αποτυγχάνουν να εκτελέσουν το έργο της εξαγωγής, λόγω των τροποποιήσεων στη δομή των σχετικών ιστοσελίδων. Το πρωτόκολλο επισκευής μπορεί να είναι επιτυχές και στην περίπτωση αυτή η εξαγωγή των δεδομένων συνεχίζεται, ή μπορεί να αποτύχει και σε αυτήν την περίπτωση προκύπτουν προειδοποιητικά μηνύματα και ανακοινώσεις. Τα σχήματα XML καθορίζονται με τη μορφή ενός DTD (Document Type Definition) και τα έγγραφα HTML παρουσιάζονται ως δέντρα DOM. Το σύστημα SG-WRAM κατασκευάζει αντίστοιχες αντιστοιχίσεις μεταξύ τους και δημιουργεί κανόνες εξαγωγής υπό την μορφή εκφράσεων XQuery.

Αυτόματη προσαρμογή wrapper. Μία άλλη στρατηγική για την αυτόματη συντήρηση των Web wrappers έχει παρουσιαστεί πρόσφατα (Ferrara & Baumgartner, 2011). Πρόκειται για μία μέθοδο αυτόματης προσαρμογής wrapper που βασίζεται στην ιδέα της σύγκρισης χρήσιμων δομικών πληροφοριών που είναι αποθηκευμένες στον Web wrapper που έχει καθοριστεί στην αρχική έκδοση της ιστοσελίδας, αναζητώντας για ομοιότητες στη νέα έκδοση της σελίδας, μετά από οποιαδήποτε δομική τροποποίηση που έχει πραγματοποιηθεί.

Η στρατηγική λειτουργεί για διαφορετικές τεχνικές εξαγωγής δεδομένων που υλοποιούνται από το σύστημα του wrapping. Για παράδειγμα, έχει δοκιμαστεί με την χρήση τόσο της γλώσσας XPath, όσο και της Elog γλώσσας wrapping. Σε αυτή τη στρατηγική, προσδιορίζονται τα στοιχεία και παρουσιάζονται ως υποδέντρα του δέντρου DOM της ιστοσελίδας και μπορούν να αξιοποιηθούν για την εύρεση των ομοιοτήτων μεταξύ δύο διαφορετικών εκδόσεων του ίδιου εγγράφου. Εξετάζουμε ένα παράδειγμα υιοθετώντας την XPath για ένα μόνο στοιχείο σε μια ιστοσελίδα.

Το σκεπτικό πίσω από την αυτόματη προσαρμογή του wrapper είναι η αναζήτηση ορισμένων στοιχείων στο τροποποιημένο κείμενο της ιστοσελίδας, που μοιράζονται δομικές ομοιότητες με την αρχική. Η αξιολόγηση της ομοιότητας γίνεται επί τη βάσει των συγκρίσιμων χαρακτηριστικών (π.χ., υποδένδρων, γνωρίσματα, κ.λπ.). Τα στοιχεία αυτά ονομάζονται υποψήφια: ανάμεσα σε αυτά, για εκείνο που δείχνει τον υψηλότερο βαθμό ομοιότητας με το στοιχείο στην αρχική σελίδα, γίνεται αντιστοίχισή του με τη νέα έκδοση της σελίδας.

Ο αλγόριθμος που υιοθετήθηκε για τον υπολογισμό της αντιστοίχισης ανάμεσα στα δέντρα DOM για τα δύο έγγραφα HTML είναι η σταθμισμένη αντιστοίχιση δέντρου. Χρησιμοποιούνται περαιτέρω heuristics για την εκτίμηση της ομοιότητας των κόμβων, για παράδειγμα, χρησιμοποιώντας τα επιπλέον χαρακτηριστικά που παρουσιάζουν τα στοιχεία DOM. Σε ορισμένες περιπτώσεις, για παράδειγμα, το περιεχόμενο του κειμένου τους θα μπορούσε να συγκριθεί, σύμφωνα με τις μετρήσεις της απόστασης των συμβολοσειρών όπως τα oJaro-Winkler (1999) ή τα δίγραμμα, ώστε να ληφθεί υπόψη και η ομοιότητα του περιεχομένου των δύο δεδομένων

κόμβων.

Είναι δυνατόν να επεκταθεί η ίδια προσέγγιση και στην περίπτωση κατά την οποία η XPath προσδιορίζει πολλαπλά παρόμοια στοιχεία στην αρχική σελίδα (π.χ., ένα XPath που επιλέγει τα αποτελέσματα μιας αναζήτησης σε ένα online κατάστημα λιανικής πώλησης, τα οποία παρουσιάζονται ως γραμμές του πίνακα, divs ή αντικείμενα λίστας. Αναλυτικότερα, είναι δυνατόν να εντοπιστούν πολλαπλά στοιχεία που μοιράζονται μία παρόμοια δομή στη νέα σελίδα, μέσα σε ένα προσαρμοσμένο επίπεδο ακρίβειας (π.χ., με την θέσπιση μίας οριακής τιμής για την ομοιότητα).

Οι συγγραφείς εφήρμοσαν την προσέγγιση αυτή σε ένα εμπορικό εργαλείο, το Lixto, περιγράφοντας το πώς η διοχέτευση της δημιουργίας wrapper έχει τροποποιηθεί ώστε να επιτρέπει στα wrappers να εντοπίζουν αυτόματα και να προσαρμόζουν την λειτουργία τους στις δομικές αλλαγές που πραγματοποιούνται στις ιστοσελίδες (Ferrara & Baumgartner, 2011). Στο πλαίσιο αυτό, η στρατηγική που προτάθηκε ήταν η απόκτηση δομικών πληροφοριών σχετικά με τα στοιχεία που εξάγει το αρχικό wrapper και την αποθήκευση τους απευθείας μέσα στο ίδιο το wrapper. Αυτό γίνεται, για παράδειγμα, με την δημιουργία υπογραφών που αντιπροσωπεύουν το υποδέντρο DOM των εξαγόμενων στοιχείων από την αρχική ιστοσελίδα, την αποθήκευσή τους ως διάγραμμα δέντρου ή ως απλά έγγραφα XML. Κατά την εκτέλεση των Web wrappers, αν προκύψει οποιοδήποτε θέμα στην εξαγωγή που οφείλεται στην δομική τροποποίηση της σελίδας, ο αλγόριθμος προσαρμογής του wrapper ξεκινά αυτόματα και προσπαθεί να προσαρμόσει το wrapper στην νέα δομή.

3.3 Υβριδικά συστήματα: δημιουργία wrapper με βάση την μάθηση

Τα συστήματα δημιουργίας wrapper και οι τεχνικές επαγωγής wrapper που συζητήθηκαν παραπάνω διαφέρουν ουσιαστικά σε δύο σημεία: στον βαθμό αυτοματοποίησης των συστημάτων Εξαγωγής Δεδομένων Ιστού και στο μέγεθος και το είδος της ανθρώπινης εμπλοκής που απαιτείται κατά τη διάρκεια της λειτουργίας.

Το πρώτο σημείο σχετίζεται με την ικανότητα του συστήματος να λειτουργήσει με έναν αυτόνομο τρόπο, διασφαλίζοντας επαρκή πρότυπα σταθερότητας και αξιοπιστίας, σύμφωνα με τις απαιτήσεις των χρηστών. Όσον αφορά στο δεύτερο σημείο, τα περισσότερα από τα συστήματα επαγωγής wrapper απαιτούν παραδείγματα με ετικέτες που παρέχονται κατά τη διάρκεια των περιόδων εκμάθησης, απαιτώντας έτσι εκ νέου την ανθρώπινη εμπλοκή ειδικών για την φάση της επισήμανσης χειρωνακτικά. Τα συστήματα δημιουργίας wrapper, από την άλλη πλευρά, εμπλέκουν τους χρήστες στην συντήρησή τους, εκτός εάν χρησιμοποιούνται αυτόματες τεχνικές, όπως αυτές που συζητήθηκαν παραπάνω.

Μία νέα κατηγορία πλατφορμών έχει εξεταστεί από την πρόσφατη βιβλιογραφία, η οποία υιοθετεί την υβριδική προσέγγιση που βρίσκεται μεταξύ της εκμάθησης με βάση τα συστήματα επαγωγής wrapper και τις πλατφόρμες δημιουργίας wrapper. Το πρώτο παράδειγμα αυτής της κατηγορίας των συστημάτων δίδεται από το RoadRunner (Crescenzi et al.,

2001), ένα σύστημα που βασίζεται στο πρότυπο το οποίο δημιουργεί τα πρότυπα για την εξαγωγή δεδομένων αυτόματα, αντιστοιχώντας τα χαρακτηριστικά από διαφορετικές σελίδες στον ίδιο τομέα. Μια άλλη ενδιαφέρουσα προσέγγιση είναι αυτή της εκμετάλλευσης των οπτικών στοιχείων και της χωρικής συλλογιστική για τον εντοπισμό στοιχείων στις ιστοσελίδες με ένα παράδειγμα προσανατολισμένο στην Υπολογιστική Όραση.

Αντιστοίχιση με βάση το πρότυπο. Το πρώτο παράδειγμα του υβριδικού συστήματος παρέχεται από το RoadRunner. Το σύστημα αυτό θα μπορούσε να θεωρηθεί ως ένα ενδιαφέρον παράδειγμα αυτόματης δημιουργίας wrapper. Η κύρια δύναμη του RoadRunner είναι ότι είναι προσανατολισμένο σε ιστοσελίδες έντασης δεδομένων που βασίζονται σε πρότυπα ή κανονικές δομές. Το σύστημα αντιμετωπίζει το πρόβλημα της εξαγωγής δεδομένων εκμεταλλευόμενο και τα δύο χαρακτηριστικά που χρησιμοποιούνται από τις γεννήτριες wrapper και από τα συστήματα επαγωγής wrapper. Πιο συγκεκριμένα, το RoadRunner μπορεί να λειτουργήσει χρησιμοποιώντας πληροφορίες που παρέχονται από τους χρήστες, με τη μορφή παραδειγμάτων επισημασμένων σελίδων ή επίσης με την αυτόματη επισήμανση ιστοσελίδων (όπως το WIEN), για την δημιουργία ενός συνόλου εκμάθησης.

Επιπλέον, θα μπορούσε να εκμεταλλεύεται την a priori γνώση σχετικά με το σχέδιο των ιστοσελίδων, για παράδειγμα, λαμβάνοντας υπόψη τα πρότυπα σελίδας που έχει μάθει από πριν. Το RoadRunner βασίζεται στην ιδέα της συνεργασίας με δύο σελίδες HTML ταυτόχρονα, προκειμένου να ανακαλύψει τα πρότυπα αναλύοντας τις ομοιότητες και τις διαφορές μεταξύ της δομής και του περιεχομένου του κάθε ζεύγους σελίδων. Ουσιαστικά, το RoadRunner μπορεί να εξάγει τις σχετικές πληροφορίες από οποιονδήποτε ιστότοπο που περιέχει τουλάχιστον δύο ιστοσελίδες με παρόμοια δομή.

Επειδή συνήθως οι ιστοσελίδες δημιουργούνται δυναμικά ξεκινώντας από το πρότυπο και τα σχετικά δεδομένα τοποθετούνται στις ίδιες (ή σε παρόμοιες) περιοχές της σελίδας, το RoadRunner είναι σε θέση να εκμεταλλευτεί αυτό το χαρακτηριστικό για τον εντοπισμό σχετικών τμημάτων πληροφοριών και ταυτόχρονα, να λαμβάνει υπόψη τις μικρές διαφορές που οφείλονται σε χαμένες τιμές ή άλλες αναντιστοιχίες.

Οι συγγραφείς όρισαν ως κατηγορία σελίδων (class of pages) εκείνες τις πηγές του Ιστού που χαρακτηρίζονται από ένα σενάριο εντολών κοινής δημιουργίας. Στη συνέχεια, το πρόβλημα ανάγεται στην εξαγωγή των σχετικών δεδομένων με τη δημιουργία wrappers για την κατηγορία σελίδων, ξεκινώντας από το συμπέρασμα μιας κοινής δομής από την σύγκριση των δύο σελίδων.

Αυτό το σύστημα μπορεί να χειριστεί χαμένες και προαιρετικές τιμές και δομικές διαφορές και προσαρμόζεται αρκετά καλά σε όλα τα είδη των πηγών του Ιστού έντασης δεδομένων. Ένα άλλο σημαντικό στοιχείο του RoadRunner είναι η υψηλής ποιότητας εφαρμογή ανοικτού κώδικα (βλέπε: www.dia.uniroma3.it/db/roadRunner/), ο οποίος παρέχει υψηλό βαθμό αξιοπιστίας του συστήματος εξαγωγής.

Χωρική συλλογιστική. Το παράδειγμα της Υπολογιστικής Όρασης

έχει εμπνεύσει επίσης το πεδίο των συστημάτων Εξαγωγής Δεδομένων Ιστού και πρόσφατα παρουσιάστηκε ένα μοντέλο εξαγωγής δεδομένων, που ονομάζεται Visual Box Model (Gatterbauer & Bohunsky, 2006).

Το Visual Box Model εκμεταλλεύεται τα οπτικά ερεθίσματα για να καταλάβει εάν στην έκδοση της ιστοσελίδας που εμφανίζεται στην οθόνη, μετά την απόδοση του προγράμματος περιήγησης υπάρχουν, για παράδειγμα, δεδομένα σε μορφή πίνακα. Το πλεονέκτημα αυτής της στρατηγικής είναι ότι είναι δυνατό να αποκτηθούν δεδομένα που δεν εμφανίζονται αναγκαστικά μέσω του προτύπου HTML <table> format.

Η λειτουργία αυτής της τεχνικής βασίζεται σε έναν αλγόριθμο OCR τύπου X-Y cut. Ο αλγόριθμος αυτός είναι σε θέση, με δεδομένη την απεικονιζόμενη έκδοση μιας ιστοσελίδας, να δημιουργήσει ένα οπτικό δίκτυο όπου τα στοιχεία της σελίδας κατανέμονται σύμφωνα με τις συντεταγμένες τους, οι οποίες καθορίζονται με οπτικές ενδείξεις. Αναδρομικά εφαρμόζονται περικοπές στην εικόνα bitmap που παρουσιάζει η απόδοση της ιστοσελίδας και αποθηκεύονται σε ένα δέντρο X-Y. Αυτό το δέντρο δημιουργείται έτσι ώστε οι προγονικοί κόμβοι με φύλλα να αντιπροσωπεύουν τους μη-κενούς πίνακες. Μερικές πρόσθετες λειτουργίες ελέγχουν αν οι εξαχθέντες πίνακες περιέχουν χρήσιμες πληροφορίες. Αυτό γίνεται διότι, αν και αποτελεί παρωχημένη πρακτική, πολλές ιστοσελίδες χρησιμοποιούν πίνακες για δομικές και γραφικές εφαρμογές, αντί για λόγους αναπαράστασης δεδομένων.

Το σύστημα εξαγωγής δεδομένων Visual Box Model υλοποιείται μέσω ενός μηχανισμού εσωτερικής απόδοσης που παράγει μία απεικόνιση της ιστοσελίδας που βασίζεται στο Gecko (developer.mozilla.org/en/Gecko), την ίδια μηχανή απόδοσης που χρησιμοποιείται από το πρόγραμμα περιήγησης Mozilla(developer.mozilla.org/it/XPCOM).

Κεφάλαιο 4ο Τεχνικές εξόρυξης ενδιαφέρουσας Πληροφορίας

4.1 Εξόρυξη από δεδομένα

Σε αυτή την ενότητα εισερχόμαστε σε λεπτομέρειες σχετικά με τα χαρακτηριστικά των υφιστάμενων συστημάτων εξαγωγής δεδομένων Ιστού. Μπορούμε να ορίσουμε γενικά ένα σύστημα εξαγωγής δεδομένων Ιστού ως πλατφόρμα εφαρμογής μιας σειράς διαδικασιών (για παράδειγμα, Web wrappers) που εξάγουν πληροφορίες από τις πηγές Ιστού (Laender et all, 2002).

Ένας μεγάλος αριθμός των συστημάτων Εξαγωγής Δεδομένων Ιστού είναι διαθέσιμος ως εμπορικά προϊόντα, ακόμη και αν ένας ολοένα και μεγαλύτερος αριθμός δωρεάν εναλλακτικών επιλογών ανοικτού κώδικα για εμπορικό λογισμικό εισάγεται πλέον στην αγορά.

Στις περισσότερες περιπτώσεις, ο μέσος όρος των τελικών χρηστών των συστημάτων Εξαγωγής Δεδομένων Ιστού είναι επιχειρήσεις ή αναλυτές δεδομένων που αναζητούν σχετικές πληροφορίες στον Ιστό. Μια ενδιάμεση κατηγορία χρηστών αποτελείται από άτομα που δεν είναι ειδικοί και που αναζητούν να συλλέξουν κάποιο περιεχόμενο Ιστού, συχνά σε μη τακτική βάση.

Αυτή η κατηγορία χρηστών είναι συχνά μη έμπειροι και ψάχνουν για απλές αλλά ισχυρές σουίτες λογισμικού Εξαγωγής Δεδομένων Ιστού, μεταξύ των οποίων μπορούμε να αναφέρουμε το DEiXTo1. Το DEiXTo βασίζεται στο W3C Document Object Model και επιτρέπει στους χρήστες να δημιουργούν εύκολα κανόνες εξαγωγής επισημαίνοντας το τμήμα των δεδομένων πάνω στο οποία θα επικεντρωθεί η εξαγωγή σε έναν ιστότοπο.

Στις επόμενες υποενότητες απεικονίζονται αρχικά οι διάφορες φάσεις που χαρακτηρίζουν ένα σύστημα Εξαγωγής Δεδομένων Ιστού και στην συνέχεια, παρουσιάζονται διάφοροι παράγοντες που επηρεάζουν τον σχεδιασμό και την εφαρμογή του συστήματος Εξαγωγής Δεδομένων Ιστού (π.χ. το έργο της δημιουργίας wrappers σύμφωνα με την ευκολία στην χρήση).

Στην συνέχεια απεικονίζεται η τεχνολογική εξέλιξη των συστημάτων Εξαγωγής Δεδομένων Ιστού σύμφωνα με κάθε παράγοντα υπό έρευνα. Για τον σκοπό αυτό, χρησιμοποιούνται ad hoc διαγράμματα που αποτελούνται από ένα σετ επιπέδων (layer cakes) έτσι ώστε τα κάτω επίπεδα (resp., top) να αντιστοιχούν στις παλαιότερες (resp., latest) τεχνολογικές λύσεις.

4.2 Οι κύριες φάσεις που σχετίζονται με το σύστημα Εξαγωγής Δεδομένων Ιστού

Σε αυτή την ενότητα περιγράφονται οι διάφορες φάσεις που σχετίζονται με την διαδικασία εξαγωγής δεδομένων από έναν ιστότοπο. Αλληλεπίδραση με ιστοσελίδες. Η πρώτη φάση ενός γενικού συστήματος Εξαγωγής Δεδομένων Ιστού είναι η αλληλεπίδραση με τον Ιστό (Wang 2000. et all): το σύστημα Εξαγωγής Δεδομένων Ιστού έχει πρόσβαση σε μια πηγή στον Ιστό και εξάγει πληροφορίες που είναι αποθηκευμένες σε αυτήν. Οι πηγές του Ιστού συμπίπτουν συνήθως με τις ιστοσελίδες, αλλά ορισμένες προσεγγίσεις λαμβάνουν επίσης υπόψη τα RSS/Atom feeds (Hammersley, 2005) και τα Microformats (Khare et all, 2006)

Μερικά εμπορικά συστήματα, π.χ. Lixto, Kapow Mashup Server, περιλαμβάνουν μια γραφική διασύνδεση χρήστη (Graphical User Interface) για την πλήρη οπτική και διαδραστική πλοήγηση των σελίδων HTML, η οποία είναι ενσωματωμένη με τα εργαλεία εξαγωγής δεδομένων.

Τα πιο προηγμένα συστήματα εξαγωγής δεδομένων Ιστού υποστηρίζουν την εξαγωγή των δεδομένων από σελίδες που έχουν προσεγγιστεί από την πλοήγηση στο λεγόμενο Βαθύ Ιστό (deep Web navigation) (Baumgartner et all, 2005): προσομοιώνουν την δραστηριότητα των χρηστών που κάνουν κλικ στα στοιχεία DOM των σελίδων, μέσα από μακροεντολές ή, πιο απλά, από την συμπλήρωση φορμών HTML.

Αυτά τα συστήματα υποστηρίζουν επίσης την εξαγωγή πληροφοριών από δυναμικές ιστοσελίδες, που συνήθως δημιουργούνται κατά τον χρόνο εκτέλεσης ως συνέπεια ενός αιτήματος χρήστη που συμπληρώνει ένα πρότυπο σελίδας με δεδομένα από κάποια βάση δεδομένων. Οι άλλες σελίδων συνήθως αποκαλούνται στατικές ιστοσελίδες, λόγω του στατικού περιεχομένου τους.

Το OxPath το οποίο αποτελεί μέρος του προγράμματος DIADEM (Furche et all. 2012) είναι ένας δηλωτικός φορμαλισμός που εκτείνεται του XPath για την στήριξη της πλοήγησης στον Βαθύ Ιστό και την εξαγωγή δεδομένων από διαδραστικές ιστοσελίδες. Προσθέτει νέα βήματα θέσης για την προσομοίωση των ενεργειών του χρήστη, έναν νέο άξονα για την επιλογή δυναμικά υπολογισμένων χαρακτηριστικών CSS, μια βολική λειτουργία πεδίου για τον εντοπισμό μόνο ορατών πεδίων, την επανάληψη των σελίδων με μια επέκταση Kleene-star για επαναλαμβανόμενες πλοηγήσεις, και νέα κατηγορήματα για την σήμανση εκφράσεων για τον εντοπισμό της εξαγωγής.

Δημιουργία ενός wrapper. Ένα σύστημα Εξαγωγής Δεδομένων Ιστού πρέπει να εφαρμόσει την υποστήριξη για την δημιουργία και την εκτέλεση wrapper.

Ένας άλλος ορισμός του συστήματος Εξαγωγής Δεδομένων Ιστού δόθηκε από τους Baumgartner et al.(2012), σύμφωνα με τους οποίους το σύστημα εξαγωγής δεδομένων Ιστού ορίζεται ως «ένα λογισμικό που εξάγει, αυτόματα και κατ' επανάληψη, τα δεδομένα από ιστοσελίδες με μεταβαλλόμενο περιεχόμενο και παραδίδει τα εξαχθέντα δεδομένα σε μια βάση δεδομένων ή σε κάποια άλλη εφαρμογή». Αυτός είναι ο ορισμός που ταιριάζει καλύτερα στην σύγχρονη αντίληψη του προβλήματος της Εξαγωγής Δεδομένων Ιστού, καθώς εισάγει τρεις σημαντικές πτυχές:

- Αυτοματοποίηση και προγραμματισμός
- Μετασχηματισμό των δεδομένων και
- Χρήση των εξαγόμενων δεδομένων

Στην συνέχεια αναλύεται κάθε μία από αυτές τις πτυχές λεπτομερώς.

Αυτοματισμός και Εξαγωγή. Η αυτοματοποίηση της πρόσβασης στις ιστοσελίδες, καθώς και στον εντοπισμό των στοιχείων τους, αποτελεί ένα από τα πιο σημαντικά χαρακτηριστικά που περιλαμβάνονται στα τελευταία συστήματα Εξαγωγής Δεδομένων Ιστού (Phan, et all, 2005).

Ανάμεσα στα πιο σημαντικά χαρακτηριστικά αυτοματισμού παραθέτουμε την δυνατότητα προσομοίωσης της ακολουθίας κλικ (clickstream) του χρήστη, την συμπλήρωση φορμών και την επιλογή μενού και κουμπιών, την υποστήριξη για την τεχνολογία AJAX (Garrett et all, 2005) για τον χειρισμό της ασύγχρονης επικαιροποίησης της σελίδας και την ικανότητα του προγραμματισμού των διαδικασιών Εξαγωγής Δεδομένων Ιστού σε περιοδική βάση.

Μετασχηματισμός των δεδομένων. Οι πληροφορίες θα μπορούσαν να προκύπτουν από πολλαπλές πηγές, πράγμα που σημαίνει χρήση διαφορετικών wrappers και επίσης, κατά πάσα πιθανότητα, απόκτηση διαφορετικών δομών των δεδομένων που εξήχθησαν. Τα βήματα μεταξύ της εξαγωγής και της παράδοσης των δεδομένων ονομάζεται μετασχηματισμός δεδομένων: κατά την διάρκεια αυτών των φάσεων, όπως καθαρισμός δεδομένων (Rahm and Do 2000) και επίλυση των συγκρούσεων (Monge, 2000), οι χρήστες επιτυγχάνουν τον στόχο της απόκτησης ομοιογενών πληροφοριών κάτω από μια μοναδική προκύπτουσα δομή. Τα πιο ισχυρά συστήματα εξαγωγής δεδομένων Ιστού

παρέχουν εργαλεία για την εκτέλεση σχήματος αυτόματης αντιστοίχισης πολλαπλά wrappers (Rahm and Bernstein, 2001) και κατόπιν τα δεδομένα συσκευάζονται στην επιθυμητή μορφή (π.χ., σε μια βάση δεδομένων, XML, κλπ), ώστε να είναι δυνατή η ανάλυση των δεδομένων, η εξομάλυνση της δομής και ο αποδιπλασιασμός των πλειάδων.

Χρήση των εξαχθέντων δεδομένων. Όταν η εργασία εξαγωγής ολοκληρωθεί και τα αποκτηθέντα δεδομένα συσκευαστούν στην απαιτούμενη μορφή, η πληροφορία αυτή είναι έτοιμη να χρησιμοποιηθεί. Το τελευταίο βήμα είναι η παράδοση του πακέτου δεδομένων, το οποίο πλέον παρουσιάζεται ως δομημένα δεδομένα, σε ένα σύστημα διαχείρισης (π.χ., μία εγγενή βάση δεδομένων XML DBMS, ένα RDBMS, μία αποθήκη δεδομένων, ένα CMS, κλπ). Επιπλέον, τα αποκτηθέντα δεδομένα μπορούν να χρησιμοποιηθούν γενικά για σκοπούς ανάλυσης ή στατιστικής (Berthold and Hand, 1999), ή απλά για να αναδημοσιευθούν σε δομημένη μορφή.

4.3 Συγκρίσεις των Layer cake

Στην ενότητα αυτή συνοψίζονται οι ικανότητες στοίβας των συστημάτων εξαγωγής δεδομένων Ιστού, συμπεριλαμβανομένων των πτυχών της δημιουργίας wrapper, τις δυνατότητες εξαγωγής δεδομένων και την χρησιμότητα του wrapper. Πιο συγκεκριμένα, εισάγονται ορισμένες ειδικές πτυχές και παρουσιάζεται η τεχνολογική εξέλιξη των συστημάτων Εξαγωγής Δεδομένων Ιστού σε σχέση με την κάθε πτυχή. Χρησιμοποιούμε ad hoc διαγράμματα δομημένα ως μια ομάδα επιπέδων (layer cakes). Σε αυτά τα διαγράμματα, τα κάτω επίπεδα (resp., top) αντιστοιχούν στις παλαιότερες (resp., latest) τεχνολογικές λύσεις.

Δημιουργία wrapper: Ευκολία Χρήσης. Οι πρώτες προσεγγίσεις για την κατανάλωση στοιχείων από τον Ιστό υλοποιήθηκαν μέσω γλωσσών γενικού σκοπού. Με την πάροδο του χρόνου, οι βιβλιοθήκες (π.χ. Ruby Mechanize) και οι γλώσσες ερωταποκρίσεων ειδικού σκοπού εξελίχθηκαν και έφτασαν στην κορυφή αυτής της αρχής (π.χ., Jedi (Huck, 1998).

Τα Wizards που απλοποιούν τον τρόπο καθορισμού των ερωταποκρίσεων είναι το επόμενο λογικό επίπεδο και, για παράδειγμα, έχουν χρησιμοποιηθεί σε W4F (Sahuguet and Azavant 1999) και XWrap (Liu et all, 2000). Τα προηγμένα συστήματα Εξαγωγής Δεδομένων Ιστού προσφέρουν γραφικές διασυνδέσεις χρήστη (GUI) για διαμόρφωση, είτε ως συστήματα με βάση τον πελάτη (π.χ. Λάπις), ή με βάση τον Ιστό (π.χ. Dapper και Needlebase) ή ως επεκτάσεις του προγράμματος περιήγησης (π.χ. iOpus και Chickenfoot). Τα εμπορικά πλαίσια προσφέρουν ένα πλήρες IDE (Ολοκληρωμένο Περιβάλλον Ανάπτυξης) με λειτουργίες π.χ. Denodo, Kapowtech, Lixto και Mozenda.

Δημιουργία wrapper: Δημιουργία Παραδείγματος. Από τη σκοπιά του πώς το σύστημα υποστηρίζει τον σχεδιαστή του wrapper στην δημιουργία ισχυρών προγραμμάτων εξαγωγής, η απλούστερη προσέγγιση είναι ο καθορισμός των ερωταποκρίσεων με το χέρι και ο έλεγχός τους σε ένα δείγμα ιστοτόπων ξεχωριστά. Οι προχωρημένοι συντάκτες προσφέρουν τονισμό των λέξεων-κλειδιών της αναζήτησης και οι χειριστές και

βοηθούν στην σύνταξη των ερωταποκρίσεων με αυτόματη συμπλήρωση και βελτιώσεις παρόμοιας χρηστικότητας (π.χ., Screen-Scraper). Στην περίπτωση των διαδικαστικών γλωσσών, τα μέσα εντοπισμού σφαλμάτων και η οπτική βοήθεια για κατασκευάσματα όπως βρόγχοι, αποτελούν πρόσθετα μέσα για την υποβοηθούμενη δημιουργία wrapper.

Όσον αφορά στον ορισμό των πλοηγήσεων στον Βαθύ Ιστό, μια σειρά από εργαλεία προσφέρουν καταγραφή τύπου βιντεοσκόπησης της ανθρώπινης περιήγησης και επανάληψη της καταγραφείσας ακολουθίας πλοήγησης (π.χ., Chickenfoot, iOpus, Lixto). Οπτικές και διαδραστικές διευκολύνσεις επίσης προσφέρονται από τα συστήματα για την απλούστευση της εξαγωγής. Οι χρήστες σημειώνουν ένα παράδειγμα περίπτωσης και το σύστημα προσδιορίζει το επιλεγμένο στοιχείο με ισχυρό τρόπο, και ενδεχομένως το γενικεύει για να ταιριάζει περαιτέρω με παρόμοια στοιχεία (π.χ., για τον εντοπισμό όλων των τίτλων βιβλίων).

Τέτοια μέσα είναι συχνά εξοπλισμένα με μεθόδους μηχανικής μάθησης, στις οποίες ο χρήστης μπορεί να επιλέξει μια πληθώρα θετικών και αρνητικών παραδειγμάτων, καθώς και το σύστημα παράγει μια γραμματική για τον εντοπισμό των αντικειμένων Ιστού υπό εξέταση, συχνά με επαναληπτική τρόπο (π.χ. Wien, Dapper, Needlebase).

Οι επιλογές «click and drag and drop» απλοποιούν περαιτέρω τους διαδραστικούς και οπτικούς μηχανισμούς. Τέλος, ορισμένα συστήματα προσφέρουν κάθετα πρότυπα για την εύκολη δημιουργία wrapper για συγκεκριμένες διευθύνσεις Ιστού, π.χ. για την εξαγωγή δεδομένων ξενοδοχείων ή στοιχείων ειδήσεων, χρησιμοποιώντας τεχνικές Επεξεργασίας Φυσικής Γλώσσας και γνώσης του πεδίου, ή για την εξαγωγή δεδομένων από τυπικά, σχέδια Ιστού, όπως δομές πινάκων ή σελίδες επισκόπησης με επόμενους συνδέσμους.

Δυνατότητες πλοήγησης στον Βαθύ Ιστό. Πριν από την έλευση των τεχνικών Web 2.0, το δυναμικό HTML και AJAX, συνήθως αρκούσε το να θεωρεί κανείς τον Ιστό ως μία συλλογή συνδεδεμένων σελίδων. Σε τέτοιες περιπτώσεις, η συμπλήρωση φορμών μπορεί να προσομοιωθεί με την παρακολούθηση των αιτήσεων και απαντήσεων από τον Διακομιστή Ιστού (Web Server) και την αναπαραγωγή της ακολουθίας των αιτήσεων (μερικές φορές με την συμπλήρωση ενός αναγνωριστικού της συνόδου δυναμικά, το οποίο έχει εξαχθεί από μια προηγούμενη σελίδα της ακολουθίας). Εναλλακτικά, τα παλαιότερα συστήματα εξαγωγής δεδομένων Ιστού έχουν επηρεαστεί από τις τεχνολογίες screen scraping σαν να χρησιμοποιούνταν για την αυτοματοποίηση 3270 εφαρμογών ή σαν να χρησιμοποιούνταν για την αυτοματοποίηση εγγενών εφαρμογών, που συνήθως βασίζονται σε μεγάλο βαθμό στις συντεταγμένες.

Η κατανόηση και η αναπαραγωγή των γεγονότων DOM στα αντικείμενα του Ιστού είναι το επόμενο λογικό επίπεδο σε αυτή την ικανότητα στοίβας.

Τα προηγμένα συστήματα μάλιστα προχωρούν ένα βήμα παραπέρα, ιδιαίτερα όταν ενσωματώνουν ένα πλήρες πρόγραμμα περιήγησης: το κλικ σε ένα στοιχείο καταγράφεται με έναν ισχυρό τρόπο και κατά τη διάρκεια της αναπαραγωγής το πρόγραμμα περιήγησης έχει ενημερωθεί για να κάνει ένα οπτικό κλικ σε ένα τέτοιο στοιχείο, παραδίδοντας τον χειρισμό

του DOM στο πρόγραμμα περιήγησης και διασφαλίζοντας ότι η ιστοσελίδα καταναλώνεται ακριβώς με τον τρόπο που ο ανθρώπινος χρήστης την καταναλώνει.

Αντίστοιχες ως προς αυτές τις επιλογές είναι οι δυνατότητες για παραμετροποίηση των ακολουθιών στον Βαθύ Ιστό, καθώς και για την χρήση τεχνικών ανίχνευσης ερωταποκρίσεων για την αυτοματοποίηση της πλοήγησης στον Βαθύ Ιστό σε άγνωστες μορφές.

Δυνατότητες Εξαγωγής Δεδομένων Ιστού. Με την πάροδο του χρόνου, έχουν συζητηθεί διάφορες προσεγγίσεις για την μοντελοποίηση μιας ιστοσελίδας. Ο απλούστερος τρόπος είναι να εργαστούμε στην ακολουθία που λαμβάνεται από τον διακομιστή του Ιστού, για παράδειγμα με τη χρήση κανονικών εκφράσεων. Σε ορισμένες περιπτώσεις, αυτό επαρκεί και ακόμη είναι και η προτιμώμενη προσέγγιση στα σενάρια μεγάλης κλίμακας για την αποφυγή δημιουργίας ενός σύνθετου και μη αποδοτικού μοντέλο. Από την άλλη πλευρά, στις σύνθετες σελίδες Web 2.0 ή σε σελίδες που δεν είναι καλοσχηματισμένες, μπορεί να είναι εξαιρετικά δύσκολη η εργασία μόνο σε επίπεδο κειμένου. Επιπλέον, τέτοια wrapper δεν διατηρούνται εύκολα και συχνά χαλάνε. Η πιο κοινή προσέγγιση είναι η εργασία επί του δέντρου DOM ή σε κάποιο άλλο είδος δομής δέντρου.

Αυτή η προσέγγιση έχει ακολουθηθεί τόσο από την ακαδημαϊκή κοινότητα, όσο και από εμπορικές προσεγγίσεις. Στις ακαδημαϊκές κοινότητες, η μελέτη της εκφραστικότητα της γλώσσας στα δέντρα και τα αυτόματα δέντρων έχουν ενδιαφέρον, και από εμπορική άποψη, είναι βολική και εύρωστη η χρήση γλωσσών όπως η XPath για τον εντοπισμό αντικειμένων στον Ιστό.

Συνήθως, δεν λαμβάνονται υπόψη μόνο τα στοιχεία ενός δέντρου DOM, αλλά και τα γεγονότα, επιτρέποντας τον καθορισμό της εξαγωγής δεδομένων και των βημάτων της πλοήγησης με την ίδια προσέγγιση.

Όμως, στον Ιστό του σήμερα, πολύ συχνά το δέντρο DOM δεν συλλαμβάνει πραγματικά την βασική δομή μιας ιστοσελίδας, όπως παρουσιάζεται στον ανθρώπινο χρήστη σε ένα πρόγραμμα περιήγησης. Ένας άνθρωπος αντιλαμβάνεται κάτι ως δομή πίνακα, ενώ το δέντρο DOM περιέχει μια λίστα στοιχείων div με απόλυτη τοποθέτηση. Επιπλέον, τα δυαδικά αντικείμενα που είναι ενσωματωμένα σε ιστοσελίδες όπως το Flash, θέτουν νέες προκλήσεις και δεν καλύπτονται με μια προσέγγιση δομής δέντρου.

Ως εκ τούτου, το screen-scraping επέστρεψε στα νέα πλαίσια της εξαγωγής δεδομένων Ιστού, χρησιμοποιώντας μεθόδους από την κατανόηση εγγράφου και την χωρική συλλογιστική, όπως τις προσεγγίσεις του έργου TamCrow(Fayzrakhmanov et all, 2011), τις χωρικές επεκτάσεις XPath (Oroe t all, 2010) του έργου ABBA (Fayzrakhmanov et all, 2010) και επεκτάσεις ως προς την απόδοση που βασίζονται στο RoadRunner για την ανίχνευση ετικετών (Crescenzi et all, 2004)

Ενσωμάτωση των προγραμμάτων Ανάλυσης και Περιήγησης. Αυτή η ικανότητα στοίβας είναι στενά συνδεδεμένη με τις αυξημένες δυνατότητες του Βαθύ Ιστού, αλλά εστιάζει στην τεχνική υλοποίηση της ενσωμάτωσης

της ανάλυσης και περιήγησης. Οι απλές προσεγγίσεις δημιουργούν το δικό τους πρόγραμμα ανάλυσης για τον εντοπισμό σχετικών ετικετών HTML, ενώ οι πιο εξελιγμένες προσεγγίσεις χρησιμοποιούν τις βιβλιοθήκες DOM χωρίς να υπάρχει σχετική προβολή του προγράμματος περιήγησης.

Λόγω του γεγονότος ότι πολλά πλαίσια Εξαγωγής Δεδομένων Ιστού υλοποιούνται σε Java, τα προγράμματα περιήγησης ειδικού σκοπού, όπως είναι το πρόγραμμα περιήγησης Swing της Java και το πρόγραμμα περιήγησης ICE και τα οποία έχουν χρησιμοποιηθεί. Οι πιο ισχυρές προσεγγίσεις είναι αυτές που ενσωματώνουν ένα πρότυπο πρόγραμμα περιήγησης όπως το IE, το Firefox ή το WebKit με βάση τα προγράμματα περιήγησης. Στην περίπτωση των εφαρμογών Java, οι διασυνδέσεις όπως η γέφυρα Java-XPCOM ή βιβλιοθήκες, όπως η JRex χρησιμοποιούνται για την ενσωμάτωση του προγράμματος περιήγησης Mozilla.

Η ενσωμάτωση ενός πλήρους προγράμματος περιήγησης δεν δίνει μόνο πρόσβαση στο μοντέλο DOM, αλλά επιπλέον και σε άλλα μοντέλα χρήσιμα για την εξαγωγή δεδομένων, συμπεριλαμβανομένου του μοντέλου CSS Box.

Μερικά εργαλεία έχουν μια διαφορετική κατεύθυνση και αντί να ενσωματώνουν ένα πρόγραμμα περιήγησης, υλοποιούνται ως επέκταση του προγράμματος περιήγησης, επιβάλλοντας κάποιους περιορισμούς και μειονεκτήματα. Αντίστοιχα με τη στοίβα του προγράμματος περιήγησης βρίσκονται οι δυνατότητες διεύρυνσης των λειτουργικών δυνατοτήτων εξαγωγής σε ανάλυση αδόμητου κειμένου αξιοποιώντας τις τεχνικές της Επεξεργασίας Φυσικής Γλώσσας.

Πολυπλοκότητα των υποστηριζόμενων ενεργειών. Τα απλά εργαλεία εξαγωγής δεδομένων προσφέρουν κοινοποίηση στον Ιστό, π.χ. αν αναφέρεται μια συγκεκριμένη λέξη. Η εξατομίκευση μιας ιστοσελίδας (π.χ., κάποια εργαλεία προσφέρουν την επιλογή αλλαγής του στυλ CSS ώστε οι πιο χρησιμοποιούμενοι σύνδεσμοι να είναι πιο εμφανείς σε μια σελίδα) είναι ένα επιπλέον επίπεδο στο layer cake.

Τα πλαίσια μαζικής επεξεργασίας προσφέρουν λειτουργίες για την αναπαραγωγή πολλών εργασιών εξαγωγής (π.χ., τρέχει μέσα από πολλές διαφορετικές τιμές σε μορφή συμπλήρωσης φορμών).

Τα προηγμένα συστήματα πάνε ένα βήμα παραπέρα και χρησιμοποιούν εξελιγμένα σχέδια εξαγωγής και τεχνικές ανωνυμίας, που εξασφαλίζουν την μη πρόκληση βλάβης στις πύλες που αποτελούν στόχο με πάρα πολλές ταυτόχρονες αιτήσεις και την παράδοση των δεδομένων σε περαιτέρω εφαρμογές όπως σε πλατφόρμες πληροφοριών για την αγορά. Η εκφραστικότητα της γλώσσας ενός wrapper, συμβάλλει επίσης στην πολυπλοκότητα των υποστηριζόμενων ενεργειών.

ΕΠΙΛΟΓΟΣ

Ολοκληρώνοντας τη παρούσα μελέτη, εξήχθη το πόρισμα ότι η έννοια της εξόρυξης στοιχείων αποτελεί την εύρεση μιας πληροφορίας με τη χρήση αλγορίθμων ομαδοποίησης και των συστημάτων βάσεων δεδομένων.

Σκοπός της εξόρυξης στοιχείων είναι τα δεδομένα που θα εξαχθούν να έχουν κατανοητή υπόσταση προς τον άνθρωπο προκειμένου μέσα από τη μελέτη αυτών να είναι σε θέση να λάβει σωστές αποφάσεις.

Η έννοια της εξόρυξης στοιχείων αποτελεί μια παραπομπή σε κάθε είδος φόρμας με πολλά στοιχεία των οποίων η επεξεργασία γενικεύεται σε κάθε είδος συστήματος υποστήριξης αποφάσεων.

Ο αληθής και ουσιαστικός σκοπός της εξόρυξης στοιχείων θεωρείται η αυτοματοποιημένη διαδικασία ανάλυσης μεγάλου όγκου στοιχείων για την εξαγωγή προτύπων άγνωστο μέχρι τότε, όπως για παράδειγμα ομάδες από εγγραφές δεδομένων

Κάτι το οποίο κυρίως εμπεριέχει τη χρήση βάσης δεδομένων μπορούν να θεωρηθούν και ως περιγραφή των στοιχείων εισαγωγής που μπορούν να χρησιμεύσουν στην εκμάθηση μηχανής και στην προγνωστική ανάλυση.

Επί παραδείγματι, η εξόρυξη στοιχείων είναι δυνατόν να προσδιορίσει αρκετά σύνολα στα στοιχεία, που δεν είναι δυνατόν να χρησιμοποιηθούν εξασφαλίζοντας πιο απόλυτα και ουσιαστικά αποτελέσματα. Παρά το γεγονός ότι η συλλογή στοιχείων καθώς και η εξήγηση των πορισμάτων δεν είναι τμήμα της εξόρυξης ανήκουν στην ανακάλυψη γνώσης ως επιπλέον βήματα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Alter, S. (2013). Work system theory: overview of core concepts, extensions, and challenges for the future. Journal of the Association for Information Systems, 72.

Baumgartner R., Ceresna M., and Ledermuller G. Deepweb navigation in web data extraction. In Proc. International Conference on Computational Intelligence for Modelling, Control and Automation, pages 698{703, Washington, DC, USA, 2005. IEEE Computer Society

Baumgartner R., Gatterbauer W., and Gottlob G.. Web data extraction system. Encyclopedia of Database Systems, pages 3465{3471, 2012

Baumgartner, R. Flesca, S. and Gottlob G. (2001). Visual web information extraction with lixto. In Proc. 27th International Conference on Very Large Data Bases, pages 119{128, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Baumgartner, R., Gatterbauer, W., & Gottlob, G., (2009), “Web data extraction system”, In Encyclopedia of Database Systems (pp. 3465-3471), Springer US

Berthold M. and D. Hand. Intelligent Data Analysis: An Introduction. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1999.

Chang, C., Kayed, M., Girgis, M., and Shaalan, K., (2006), “A Survey of Web Information Extraction Systems”, IEEE Transactions on Knowledge and Data Engineering, 18(10):1411{1428

Crescenzi V., Mecca G., and Merialdo P.. Improving the expressiveness of roadrunner. In SEBD, pages 62{69, 2004.

Dalvi, N., Bohannon, P., & Sha, F., (2009), “Robust web extraction: an approach based on a probabilistic tree-edit model”, In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (pp. 335-348), ACM.

Dalvi, N., Kumar, R., & Soliman, M., (2011), “Automatic wrappers for large scale web extraction”, Proceedings of the VLDB Endowment, 4(4), 219-230

Doan, A., Naughton, J. F., Ramakrishnan, R., Baid, A., Chai, X., Chen, F., ... & Vuong, B. Q., (2009), “Information extraction challenges in managing unstructured data”, ACM SIGMOD Record, 37(4), 14-20

Durbin A., Essentials of Marketing , 14th edition , South – Western College Publishing 1997 , pp., 411-437.

Eikvil, L., (1999), “Information extraction from world wide web-a survey”

Fayzrakhmanov R., Goebel M., Holzinger W., Kruepl B., Mager A., and Baumgartner R.. Modelling web navigation with the user in mind. In Proc. 7th International Cross-Disciplinary Conference on Web Accessibility, 2010

Ferrara E. and Baumgartner R. (2011). Intelligent self-repairable web wrappers. Lecture Notes in Computer Science, 6934:274{285

Fiumara, G., (2007), “Automated information extraction from web sources: a survey”, Proc. of Between Ontologies and Folksonomies Workshop, pages 1{9

Freitag D. (2000) Machine learning for information extraction in informal domains. Machine learning, 39(2):169{202

Furche T., Gottlob G., Grasso G., Gunes O., Guo X., Kravchenko A., Orsi G., Schallhart C., Sellers A. J., and Wang C.. DIADEM: domain-centric, intelligent, automated data extraction methodology. In WWW (Companion Volume), pages 267{270, 2012.

Galliers, R. D., & Leidner, D. E. (Eds.). (2014). Strategic information management: challenges and strategies in managing information systems. Routledge.

Garrett J.. AJAX: A New Approach to Web Applications. Technical report, Adaptive Path, 2005.

Gatterbauer W. and Bohunsky P. (2006). Table extraction using spatial reasoning on the css2 visual box model. In Proc. 21st national conference on Arti_cial intelligence, pages 1313{1318. AAAI Press

Gottlob G. and Koch. C. (2004) Monadic datalog and the expressive power of languages for web information extraction. J. ACM, 51(1):74{113

Hammersley B.. Developing feeds with rss and atom. O'Reilly, 2005

Hsu C.-N. and Dung M.-T. (1998). Generating _nite-state transducers for semi-structured data extraction from the web. Inf. Syst., 23(9):521{538, 1998.

Jung, K., In Kim, K., & K Jain, A., (2004), "Text information extraction in images and video: a survey", Pattern recognition, 37(5), 977-997

Kantardzic, M. (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons

Kennedy D., " Who' s on line ;" Inc Technology, No 1 (1997) pp., 34-39

Khare R. and Celik T.. Microformats: a pragmatic path to the semantic web. In Proc. 15th international conference on World Wide Web, pages 865{866, New York, NY, USA, 2006. ACM

Kushmerick, N., (1997), "Wrapper induction for information extraction", (Doctoral dissertation, University of Washington)

Kushmerick, N., (2002), "Finite-state approaches to web information extraction", Proc. of 3rd Summer Convention on Information Extraction, pages 77{91

Kushmerick. N. (2000). Wrapper induction: e_ciency and expressiveness. Arti_cial Intelligence, 118(1-2):15{68

Laender A. H. F., B. A. Ribeiro-Neto, A. S. da Silva, and J. S. Teixeira. A brief survey of web data extraction tools. SIGMOD Rec., 31(2):84{93, 2002.

Liu L. C. Pu, and W. Han. XWrap: An extensible wrapper construction system for internet information. In Proc. 16thInternational Conference on Data Engineering, 2000.

Meng, X. Hu, D. and Li. C. (2003). Schema-guided wrapper maintenance for web-data extraction. In Proc. of the ACM international workshop on Web information and data management, pages 1{8, New York, NY, USA. ACM

Monge A. E. Matching algorithm within a duplicate detection system. IEEE Techn. Bulletin Data Engineering, 23(4), 2000.

Mooney R. (1999). Relational learning of pattern-match rules for information extraction. In Proc. of the National Conference on Arti_cial Intelligence (AAAI 1999), pages 328{333

Muslea, I. Minton, S. and Knoblock C. (1999). A hierarchical approach to wrapper induction. In Proc. of the 3rd annual conference on Autonomous Agents, pages 190{197, New York, NY, USAACM.

Oro E., Ru_olo M., and Staab. Sxpath S. - extending xpath towards spatial querying on web documents. PVLDB, 4(2):129{140, 2010.

Paul, P. K. (2014). Information Systems and Different Domain, Functionalities and Types: A Conceptual Study. pinnacle mathematics & computer science.

Phan X. H., Horiguchi S., and Ho T.. Automated data extraction from the web with conditional models. Int. J. Bus. Intell. Data Min., 1(2):194{209, 2005

Pinedo, M. L. (2012). Scheduling: theory, algorithms, and systems. Springer Science & Business Media.

Qureshi, M. S. (2012). Measuring Efficacy of Information Security Policies: A Case Study of UAE based company.

Rahm E. and Bernstein P. A survey of approaches to automatic schema matching. The VLDB Journal, 10(4):334{350,2001

Rahm E. and Do H. H.. Data cleaning: Problems and current approaches. IEEE Bulletin on Data Engineering, 23(4), 2000.

Sahuguet A. and Azavant F.. Building light-weight wrappers for legacy web data-sources using w4f. In Proc. 25th International Conference on Very Large Data Bases, pages 738{741, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc

Sarawagi, S., (2008), “Information extraction”, Foundations and trends in databases, 1(3), 261-377

Soderland S. (1999). Learning information extraction rules for semi-structured and free text. Machine learning, 34(1):233{272

Sommestad, T., Ekstedt, M., Holm, H., & Afzal, M. (2011). Security mistakes in information system deployment projects. Information Management & Computer Security, 19(2), 80-94.

Sousa, K., & Oz, E. (2014). Management information systems. Cengage Learning.

Susanto, H., Almunawar, M. N., & Tuan, Y. C. (2011). Information security management system standards: A comparative study of the big five.

Thimm, H., & Rasmussen, K. B. (2013). Obtaining informed ness in collaborative networks through automated information provisioning–a modelling framework and active database system approach. International Journal of Computer Integrated Manufacturing, 26(11), 1054-1065.

Tsvetkov, V. Y. (2014). Moscow State Technical University of Radio Engineering, Electronics and Automation MSTU MIREA, Vernadsky Prospekt, 78, Moscow, 119454, Russia Abstract. The theme of the present article concerns the information field. It is stated here that information field may be discrete or continuous. Information field includes two types: natural information field and artificial information. Life Science Journal, 11(5).

Wang P., W. Hawk, and C. Tenopir. Users' interaction with world wide web resources: an exploratory study using a holistic approach. Inf. Process. Manage., 36:229{251, January 2000.

Wang, J., Gupta, M., & Raj, R. (2015). Insider Threats in a Financial Institution: Analysis of Attack-Proneness of Information Systems Applications. Management Information Systems Quarterly, 39(1), 91-112.

Wang, J., Gupta, M., & Raj, R. (2015). Insider Threats in a Financial Institution: Analysis of Attack-Proneness of Information Systems Applications. Management Information Systems Quarterly, 39(1), 91-112.

Xu, J., Chau, M., and Tan, C. Y. B., (2014), “The Development of Social Capital in the Collaboration Network of Information Systems Scholars”, Journa of the Association of Informational Systems, vol. 15, Issue 12, pp. 835-859

Yan, L., & Ma, Z. M. (2014). Modeling fuzzy information in fuzzy extended entity-relationship model and fuzzy relational databases. Journal of Intelligent and Fuzzy Systems, 27(4), 1881-1896.

Zafar, H. (2013). Human resource information systems: Information security concerns for organizations. Human Resource Management Review, 23(1), 105-113.

Atzeni, P., Bugiotti, F., & Rossi, L. (2012, January). Uniform access to non-relational database systems: The SOS platform. In Advanced Information Systems Engineering (pp. 160-174). Springer Berlin Heidelberg.