Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ Σ Τ Α Τ Ι Σ Τ Ι Κ Η Σ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ Ε Π Ι Σ Τ Η Μ Η Σ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

Εξόρυξη Δεδομένων σε δεδομένα διαδικτυακής κίνησης και ροής χτυπημάτων

Δημήτρης Καλτσάς

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Σεπτέμβριος 2015

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. 1η /26.09.2013 συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Επίκουρος Καθηγητής Ν. Πελέκης (Επιβλέπων)
- Καθηγητής Ι. Θεοδωρίδης
- Επίκουρος Καθηγητής Ελ. Κοφίδης

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

U N I V E R S I T Y O F P I R A E U S

DEPARTMENT OF STATISTICS
AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

Data Mining in web traffic & stream clicks

By
KALTSAS DIMITRIOS

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
September, 2015

Περίληψη

Ο παγκόσμιος ιστός (www – world wide web) και το διαδίκτυο (Internet) πάνω στο οποίο στηρίζεται έφεραν την τρίτη επανάσταση του εικοστού αιώνα στον τρόπο επικοινωνίας μεταξύ των ανθρώπων. Ώρα με την ώρα γίνεται πιο δημοφιλής τρόπος επικοινωνίας και από περισσότερους χρήστες. Ένα δε σημαντικό κομμάτι της ανθρώπινης γνώσης έχει ήδη εγκατασταθεί στον παγκόσμιο ιστό. Περάσαμε λοιπόν από τα απλά δεδομένα στα "μεγάλα δεδομένα" (big data) που υπάρχουν διαθέσιμα και μας περιμένουν να εφαρμόσουμε πάνω τους διαδικασίες επεξεργασίας και στατιστικής για εξαγωγή συμπερασμάτων. Ένας κλάδος της πληροφορικής επιστήμης που ασχολείται με την επεξεργασία των δεδομένων στον παγκόσμιο ιστό χρησιμοποιεί μεθόδους εξόρυξης γνώσης και ονομάζεται "εξόρυξη παγκόσμιου ιστού".

Στην εργασία μας αυτή σκοπός ήταν να κάνουμε "εξόρυξη χρήσης παγκόσμιου ιστού". Επεξεργαστήκαμε λοιπόν τα αρχεία καταγραφής του εξυπηρετητή μιας ασφαλιστικής εταιρείας σχετικά με την επισκεψιμότητα, με σκοπό την εξαγωγή συμπερασμάτων σχετικά με τις ημερομηνίες. Συγκεκριμένα εφαρμόστηκαν αλγόριθμοι ομαδοποίησης και ταξινόμησης για την εξόρυξη δεδομένων από την επισκεψιμότητα της ιστοσελίδας. Η πληροφορία αυτή είναι χρήσιμη για τη χάραξη διαφημιστικής στρατηγικής, για το ποιες περιόδους είναι κατάλληλες, καθώς το κοινό παρουσιάζει μεγαλύτερο ενδιαφέρον.

Λέξεις κλειδιά: εξόρυξη χρήσης παγκόσμιου ιστού, εξόρυξη παγκόσμιου ιστού, εξόρυξη γνώσης, αρχεία καταγραφής εξυπηρετητή, συσταδοποίηση, κατηγοριοποίηση

Summary

The world wide web and the Internet in which it is based have brought the third revolution of the twentieth century in the way people communicate. Hour by hour it becomes the most popular means of human communication. An important piece of human knowledge has been already installed in the web. We have passed from data to “big data”. There is a lot of data in the web just waiting to be statistically analyzed and mined. A computer science field that deals with processing web data using data mining techniques is called “web mining”.

Our purpose in this thesis is to apply “web usage mining” to web data coming from web server log files. The target was an insurance company in which we had to infer results about its visitors. We also use SPSS, applying clustering techniques for data mining for the visitors. This information is useful, in order to organize the advertising strategy plan in suitable periods.

Keywords: web usage mining, web mining, data mining,, clustering, classification.

Περιεχόμενα

Summary 5
Εισαγωγή 8
ΚΕΦΑΛΑΙΟ 1 10
Εισαγωγή στην Εξόρυξη Δεδομένων από το Διαδίκτυο 10
1.1 Η Εξόρυξη Δεδομένων 10
1.2 Τα Χαρακτηριστικά Λειτουργίας της Εξόρυξης Δεδομένων 14
1.2.1 Μέθοδος Web Content Mining 15
1.2.2 Μέθοδος του Web Structure Mining 16
1.2.3 Μέθοδος του Web Usage Mining 16
1.3 Λόγοι για τους Οποίους Χρησιμοποιείται η Εξόρυξη Δεδομένων σε Διαδικτυακή Βάση 17
1.3.1 Χρήση στο Παγκόσμιο Ιστό 18
1.3.2 Χρήση σε Διάφορες Οικονομικές και Λοιπές Επιστήμες 18
1.3.3 Χρήση στο Τομέα του Μάρκετινγκ 18
1.3.4 Χρήση για Πρόβλεψη Επενδύσεων 19
1.3.5 Χρήση για Πρόληψη και Ασφάλεια 19
1.4 Στάδια Διαδικασίας 20
1.4.1 Συλλογή δεδομένων 20
1.4.2 Προεπεξεργασία δεδομένων 22
1.4.3 Ανακάλυψη προτύπων 23
1.4.4 Εκμετάλλευση της γνώσης 24
1.5 Οφέλη που Προκύπτουν Αναφορικά με την Ανάλυση Στοιχείων για Χρήστες του Διαδικτύου και το Ενδιαφέρον τους για τα Προϊόντα – Υπηρεσίες των Διαφημιζόμενων Επιχειρήσεων 24
1.6 Παράγοντες του Διαδικτύου που Επηρεάζουν την Εξόρυξη Δεδομένων 28
ΚΕΦΑΛΑΙΟ 2 30
Τεχνικές ομαδοποίησης σε διαδικτυακά δεδομένα 30
2.1 Εισαγωγή 30
2.2 Η διαδικασία του Clustering 31
2.3 Επιλογή του Κατάλληλου Αλγόριθμου 33
2.4 Ορισμοί και Συμβολισμός 34
2.5 Αναπαράσταση στοιχείων, επιλογή και εξαγωγή χαρακτηριστικών 34
2.6 Μέτρο ομοιότητας 36
2.7 Τεχνικές Clustering 37
2.8 Expectation Maximization (EM) 39
2.9 Ο αλγόριθμος k-means 41
2.10 Ιεραρχικοί αλγόριθμοι ομαδοποίησης 43
2.11 Ομαδοποίηση βασισμένη-στην-πυκνότητα (Density-based clustering) 46
ΚΕΦΑΛΑΙΟ 3 49
Εμπειρικό Μέρος 49
3.1 Περιγραφική των Δεδομένων 49
3.2 Διαφοροποιήσεις επισκέψεων στην ιστοσελίδα ανά περιοχή 51
3.3 Μεταβολές επισκεψιμότητας ανά περιοχή 54
3.4 Επίτευξη στόχων – μετατροπών ανά περιοχή 57
3.5 Πρόβλεψη του CVR με την εφαρμογή παλινδρόμησης 61
3.6 Συσταδοποίηση 66
3.6.1 Σενάριο 1: Συσταδοποίηση ημερών με κριτήριο το πλήθων των επισκεπτών/ημέρα 67

3.6.2 Σενάριο 2: Συσταδοποίηση ημερών με κριτήριο το πλήθων των νέων επισκεπτών/ημέρα....67
3.6.3 Σενάριο 3: Συσταδοποίηση ημερών με κριτήριο το μέσο αριθμό στόχων που έχει ο ιστότοπος (μετατροπών)....68
3.6.4 Σενάριο 4: Πολυκριτηριακή Συσταδοποίηση ημερών....68
Συμπεράσματα – Προτάσεις....73
Βιβλιογραφία....75

Εισαγωγή

Αποτελεί γεγονός πως μια βάση δεδομένων είναι μια συλλογή από δεδομένα. Αντίθετα με ένα απλό σύνολο, τα δεδομένα σε μια βάση έχουν μια ορισμένη δομή ή σχήμα με το οποίο είναι σχετιζόμενα. Έτσι τα δεδομένα σε μια βάση αναπαρίστανται με ένα πιο θεωρητικό τρόπο ή μοντέλο δεδομένων. Αυτό το μοντέλο χρησιμοποιείται για να περιγράψει τα δεδομένα, τα χαρακτηριστικά τους, και τις σχέσεις μεταξύ τους (Ramesh et al, 2001).

Ένα μεγάλο μέρος των σημερινών ερευνητών στην εξόρυξη δεδομένων είναι άτομα προερχόμενα από τον τομέα των βάσεων δεδομένων. Η σχέση των δύο αυτών τομέων είναι εμφανής μιας και πριν επεξεργαστούμε τα δεδομένα μας πρέπει πρώτα να μπορούμε να τα διαχειριστούμε ορθά. Έτσι χωρίς καλά συστήματα διαχείρισης δεδομένων είναι πιο δύσκολο να εφαρμόσουμε αλγόριθμους εξόρυξης δεδομένων (Berry, Linoff, 2000).

Όπως σημειώνεται λοιπόν και παρακάτω, ιδιαίτερα σημαντικές είναι οι τεχνικές εξόρυξης και ροής δεδομένων, οι οποίες βοηθούν τους ειδικούς να εξάγουν τα αποτελέσματα τα οποία επιθυμούν για μια συγκεκριμένη ομάδα ερευνών. Ως εκ τούτου και η παρούσα μεταπτυχιακή εργασία, αναφέρεται στην εξόρυξη δεδομένων σε δεδομένα διαδικτυακής βάσης, στις web data mining τεχνικές εξόρυξης δεδομένων σε διαδικτυακή βάση και τέλος στον τρόπο και χαρακτηριστικά εξόρυξης δεδομένων από διαδικτυακή βάση σύμφωνα με την ροή «χτυπημάτων».

Σκοπός μας στην εργασία αυτή είναι να εξορύξουμε γνώση (mine data) από αρχεία καταγραφής (web logs) του διαδικτυακού εξυπηρετητή (web server) της ιστοσελίδας μιας ασφαλιστικής εταιρείας (Internet portal) με υψηλή επισκεψιμότητα. Η διαδικασία αυτή που είναι γνωστή και ως εξόρυξη γνώσης διαδικτυακής χρήσης (web usage mining) βασίζεται στην εκμετάλλευση των αρχείων καταγραφής που διατηρούν οι διαδικτυακοί εξυπηρετητές. Τα τελικά αποτελέσματα θα μας δώσουν στατιστικά στοιχεία για την πλοηγική συμπεριφορά των επισκεπτών της δικτυακής πύλης και με βάση αυτά θα μπορούμε κατά κάποιο τρόπο να προβλέψουμε και τη μελλοντική συμπεριφορά τους ώστε να γίνουν κατάλληλες αλλαγές στη μορφή των σελίδων της πύλης. Οι αλλαγές αυτές συνήθως σχετίζονται με τη βελτίωση της δομής των περιεχομένων της ώστε να είναι πιο προσβάσιμα απ' ότι είναι τώρα. Επίσης

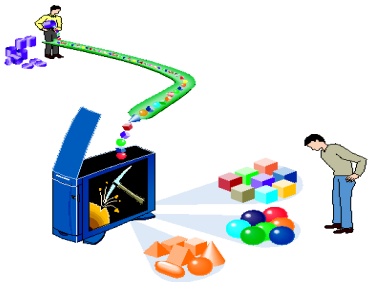
μπορεί να αφορούν την καλύτερη εμπορική εκμετάλλευση της πύλης τοποθετώντας διαφημίσεις σε κάθε σελίδα που να έχουν σχέση με τις προτιμήσεις του εκάστοτε χρήστη. Τέλος θα μπορούσαν να χρησιμοποιηθούν για την προσωποποίηση των σελίδων της πύλης και των μενού της για κάθε χρήστη ώστε να αντικατοπτρίζουν τις προτιμήσεις του.

Η προσπάθεια που έγινε στην εργασία αυτή ήταν να προσφέρει μια συνολική και τελική λύση για το σύνολο των σταδίων της εξόρυξης δεδομένων. Τα τρία αυτά στάδια αποτελούνται από την προεπεξεργασία (καθαρισμό και ανάλυση των δεδομένων των αρχείων καταγραφής), την κύρια επεξεργασία εφαρμόζοντας στατιστικές τεχνικές και τεχνικές εξόρυξης γνώσης και την έκδοση των τελικών αποτελεσμάτων με την έκδοση στατιστικών πινάκων και διαγραμμάτων.

Χρησιμοποιήσαμε λοιπόν το στατιστικό πακέτο SPSS, το οποίο παρέχει τη δυνατότητα κατηγοριοποίησης και συσταδοποίησης των δεδομένων, χωρίς την εφαρμογή προγραμματισμού.

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή στην Εξόρυξη Δεδομένων από το Διαδίκτυο



1.1 Η Εξόρυξη Δεδομένων

Αναφερόμενοι σχετικά στην Εξόρυξη Δεδομένων από το Διαδίκτυο ή διαφορετικά στην διεθνή ορολογία Web Data Mining, θα σημειώναμε πρώτιστα πως το Διαδίκτυο ξεκίνησε ως ένα ιδιαίτερα δημοφιλές «εργαλείο» στην επιστημονική κοινότητα, ενώ πλέον ο ρυθμός ανάπτυξής του είναι τέτοιος που έχει εκτιμηθεί ότι ο πληθυσμός των χρηστών του σήμερα, περίπου αγγίζει τα 2,5 δισεκατομμύρια, σημειώνοντας αύξηση της τάξης του 150% τα τελευταία 5 χρόνια (Data Source: Internet World Stats).

Η ταχεία ανάπτυξη του διαδικτύου λοιπόν, το καθιστά ως τη μεγαλύτερη και προσβάσιμη πηγή εξόρυξης δεδομένων σε παγκόσμιο επίπεδο. Τα χαρακτηριστικά του Ιστού είναι τέτοια, τα οποία οριοθετούν την Εξόρυξη Δεδομένων σε ένα πολύ συναρπαστικό μα και παράλληλα δύσκολο έργο. Για το λόγο αυτό, τα κυριότερα από αυτά τα χαρακτηριστικά αναφέρονται ως εξής (Zanasi, 2003):

- Η ποσότητα των δεδομένων - πληροφοριών είναι πραγματικά τεράστια και ανανεώνεται συνεχώς, καθιστώντας έτσι εφικτή τη δυνατότητα για κάποιο χρήστη να βρει οποιαδήποτε πληροφορία επιθυμεί.
- Εκτός από την ποσότητα, μπορεί να βρεθεί και αρχείο οποιασδήποτε ποιότητας και μορφής (κείμενο, φωτογραφία/εικόνα, video).
- Το σύνολο της πληροφόρησης που προσφέρει το Διαδίκτυο, είναι ετερογενές. Στην πράξη αυτό σημαίνει ότι μπορεί να υπάρξει η ίδια πληροφορία με

διαφορετικό περιεχόμενο, κάτι το οποίο καθιστά την ενσωμάτωση της πληροφορίας ως ένα ζήτημα δύσκολο και περίπλοκο.

- Οι περισσότερες πληροφορίες που είναι διαθέσιμες, συνδέονται μεταξύ τους με υπερσυνδέσεις (hyperlinks).
- Το σύνολο της πληροφόρησης είναι δυναμικό, αλλάζει και ανανεώνεται δηλαδή συνέχεια.
- Το διαδίκτυο δίνει σχεδόν στον κάθε ένα χρήστη του τη δυνατότητα να δίνει πληροφορίες, χωρίς να μπορεί να ελέγχει την ποιότητα της πληροφορίας αυτής.
- Εκτός από την πληροφόρηση, τα τελευταία χρόνια έχει ανθήσει και η παροχή υπηρεσιών από το Internet. Τέτοιου είδους υπηρεσίες είναι οι πωλήσεις είτε αγαθών είτε υπηρεσιών (E- Commerce)
- Τέλος, το Internet μπορεί να χαρακτηρισθεί σαν μια παγκόσμια κοινωνία, όπου υπάρχουν συνεχείς αλληλεπιδράσεις μεταξύ ατόμων είτε με άλλα άτομα, είτε με οργανισμούς.

Το διαδίκτυο λοιπόν μοιάζει ουσιαστικά μ' έναν οργανισμό ο οποίος αποτελείται από εκατομμύρια κύτταρα που είναι όλα μεταξύ τους συνδεδεμένα και ταυτόχρονα συνυφασμένα με την έννοια της επικοινωνίας. Συνεχίζοντας τον σχετικό παραλληλισμό, μπορούμε να υποστηρίξουμε πως πρόκειται για έναν τεράστιο οργανισμό που συνεχώς που εξελίσσεται και μεταλλάσσεται. Από τη δημιουργία του και την αρχική του χρήση όπως αυτή έχει περιγραφεί παραπάνω, έχει πλέον μεταλλαχθεί σε ένα δίκτυο όπου τα εκατομμύρια καταναλωτών συνδέονται με κάθε εταιρεία στον κόσμο και κατά συνέπεια με τις βάσεις δεδομένων απογραφής των προϊόντων.

Από τα παραπάνω στοιχεία, γίνεται εύκολα αντιληπτό πως μόνο ένα μέρος από το σύνολο της πληροφορίας που παρέχεται είναι αξιοποιήσιμο, με το υπόλοιπο να θεωρείται θορυβώδες. Αυτό σε συνάρτηση με το ότι όλο περισσότερο πληθαίνουν οι φωνές οι οποίες ζητούν την προστασία των προσωπικών στοιχείων των χρηστών από τους ιστοτόπους τους οποίους επισκέπτονται, καθιστούν αναγκαία τη χρήση Τεχνικών Εξόρυξης Δεδομένων (Data Mining). Η χρήση Τεχνικών Εξόρυξης Δεδομένων στο Διαδίκτυο, ονομάζεται Web Mining, και σκοπός του είναι η ανακάλυψη τρόπων χρήσης από τα δεδομένα του Web προκειμένου οι εταιρείες να

μπορέσουν να κατανοήσουν τις συμπεριφορές των πελατών τους και των δυνητικών πελατών τους (Zanasi, 2003).

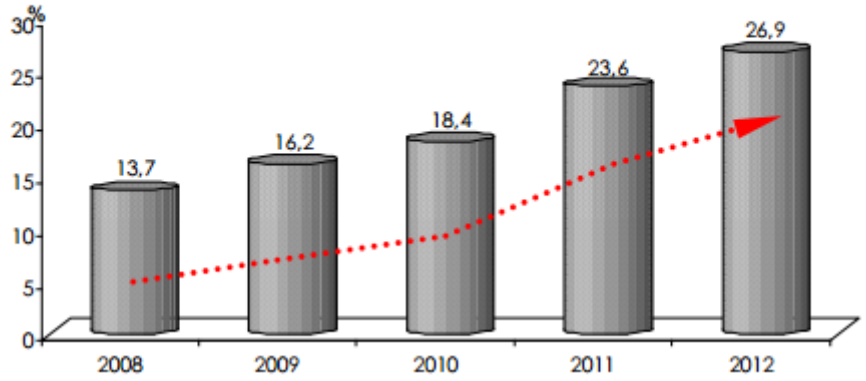
Ωστόσο θα πρέπει να σημειωθεί πως οι γνώσεις σχετικά με τους επισκέπτες και τους πελάτες μιας ιστοσελίδας ηλεκτρονικού εμπορίου, τη συμπεριφορά και τις ανάγκες αυτών, είναι απαραίτητες αφενός για τη μετατροπή του επισκέπτη σε πελάτη και αφετέρου για τη διατήρηση του πελάτη (customer loyalty) καθώς όπως αναφέρουν οι ειδικοί στις επιχειρήσεις, οι ανταγωνιστές είναι ένα κλικ μακριά. Συνεπώς προκειμένου να διατηρηθεί η επιτυχία για επιχείρηση ηλεκτρονικού εμπορίου (e-commerce) είναι αναγκαία η συλλογή και η ανάλυση στοιχείων συμπεριφορών, όσο και δημογραφικών για τους πελάτες της (Berry, Linoff, 2001). Η διαδικασία της Εξόρυξης Δεδομένων από το Διαδίκτυο ή Web Mining μπορεί να βοηθήσει μία επιχείρηση ηλεκτρονικού εμπορίου στους τομείς (Banerjee, 2001):

- Βελτίωση των πωλήσεων,
- Διαδικασία cross selling,
- Προσωποποιημένες διαφημίσεις (personalized ads)
- Click Through Rate (αναπτύσσεται σε άλλο σημείο)

αναλύοντας τα Clickstream δεδομένα της αγοράς, των επισκεπτών, των αγοραστών και συνολικά της διαδρομής του χρήστη μέχρι την ευχαριστήρια σελίδα (Thank You Page). Η αναγκαιότητα για χρήση της μεθόδου Web Mining, διαφαίνεται και από την πρόβλεψη ότι οι Ειδικοί Εξόρυξης Δεδομένων από το διαδίκτυο - Web Mining Specialists στα επόμενα 2-3 χρόνια θα χαρακτηρίζονται ως δυσεύρετοι (3rd annual Travel@Google Event, Athens March 2014).

Καταλήγοντας στη παρούσα ενότητα, θα λέγαμε λοιπόν πως η εξόρυξη δεδομένων από το διαδίκτυο, συνδέεται άμεσα με το γεγονός πως ο ανταγωνισμός μεταξύ των εταιρειών του ηλεκτρονικού εμπορίου είναι πλέον τεράστιος. Τα πλέον πρόσφατα στοιχεία αναφέρονται στο 2012 (συνεπώς οι αριθμοί θα διαφέρουν προς τα πάνω). Η Ελληνική Στατιστική Αρχή πραγματοποίησε την «Έρευνα Χρήσης Τεχνολογιών Πληροφόρησης και Επικοινωνίας από τα Νοικοκυριά για το Έτος 2012», όπου ενδεικτικά παρατηρείται ότι:

Γράφημα 1. Ηλεκτρονικό εμπόριο: Α' τρίμηνο 2008 – 2012



Πηγή: http://www.statistics.gr/

- Το ποσοστό των χρηστών του διαδικτύου που κατά το Α' τρίμηνο του 2012 πραγματοποίησαν ηλεκτρονικές αγορές ανέρχεται σε 26,9%, το οποίο υποδηλοί αύξηση κατά 14% σε σχέση με το αντίστοιχο χρονικό διάστημα του 2011.
- Συνεπεία της στροφής στην προτίμηση των καταναλωτών στις υπηρεσίες του ηλεκτρονικού εμπορίου είναι λογικό να αυξάνεται και ο ανταγωνισμός μεταξύ αυτών.
- Ομοίως με το παραδοσιακό εμπόριο, η διατήρηση των πελατών και η αύξηση των πωλήσεων είναι μονόδρομος προκειμένου τα καταστήματα του ηλεκτρονικού εμπορίου να μπορέσουν να επιβιώσουν. Τα καταστήματα οφείλουν να γνωρίζουν και να κατανοήσουν τις ανάγκες και τις προτιμήσεις των πελατών τους. Αυτή η κατανόηση απαιτεί τη γνώση σχετικά με τις προτιμήσεις, την ηλικιακή ομάδα, το φύλλο, τη συμπεριφορά και εν γένει με τον τρόπο ζωής του πελάτη.
- Κάθε ιστότοπος όμως οφείλει να σέβεται την ανωνυμία των επισκεπτών του, έτσι δεν είναι τόσο εύκολη η άντληση των δημογραφικών στοιχείων αυτών. Ακριβώς για το λόγο αυτό χρησιμοποιούν τεχνικές Εξόρυξης Δεδομένων - Data Mining για την ανάλυση των προφίλ και την εξαγωγή trends για τους επισκέπτες τους.

1.2 Τα Χαρακτηριστικά Λειτουργίας της Εξόρυξης Δεδομένων

Αποτελεί γεγονός πως από τα πλέον ενδιαφέροντα ερευνητικά πεδία του γενικού τομέα εξόρυξης δεδομένων, είναι η εξόρυξη δεδομένων στον παγκόσμιο ιστό. Όπως είναι γνωστό στους ειδικούς, το να υπολογίσουν το ακριβές μέγεθος δεδομένων του παγκόσμιου ιστού, είναι ουσιαστικά αδύνατο. Το έτος 2012 έχει υπολογιστεί πως υπάρχουν περίπου 104,854,877 σελίδες με ρυθμό αύξησης τεσσάρων περίπου εκατομμυρίων σελίδων το μήνα (Heer, 2004). Η δημοφιλής μηχανή αναζήτησης Yahoo ανακοίνωσε πρόσφατα μέσα από την σελίδα της πως έχει στο ευρετήριο της περίπου 20 εκατομμύρια αντικείμενα από τα οποία τα 19 εκατομμύρια είναι δεδομένα κειμένου. Ο παγκόσμιος ιστός μπορεί να θεωρηθεί ως η μεγαλύτερη βάση δεδομένων που είναι ανοικτή και διαθέσιμη σε κάθε χρήστη και καθημερινά αντιμετωπίζει τις προκλήσεις τόσο σε θέματα παρουσίασης όσο και ποιότητας δεδομένων (Nanopoulos et al, 2011).

Ο όρος βέβαια βάση δεδομένων χρησιμοποιείται περισσότερο θεωρητικά, μιας και στην πραγματικότητα δεν υπάρχει πρακτικά δομή ή σχήμα στον παγκόσμιο ιστό. Αυτό το γεγονός κάνει ακόμα πιο επιτακτική την ανάγκη για εξόρυξη δεδομένων στον παγκόσμιο ιστό παρέχοντας τεράστια βοήθεια σε κάθε είδους χρήστη. Με τον όρο εξόρυξη γνώσης στον παγκόσμιο ιστό, δεν αναφερόμαστε μόνο σε δεδομένα που περιέχονται σε ιστοσελίδες αλλά και σε δεδομένα που έχουν να κάνουν με τη δραστηριότητα ενός χρήστη σε αυτό (Nanopoulos et al, 2011). Τα δεδομένα διαδικτύου μπορούν να χωριστούν στις ακόλουθες κατηγορίες:

- Περιεχόμενο ιστοσελίδων
- Ενδοπληροφορία ιστοσελίδων (HTML / XML κώδικας)
- Εσωτερική δομή ιστοσελίδων, δηλαδή των πως διασυνδέονται μεταξύ τους
- Δεδομένα χρήσης που περιγράφουν πως οι επισκέπτες προσπελαύνουν τις ιστοσελίδες
- Προφίλ χρηστών που περιλαμβάνουν δημογραφικά δεδομένα και πληροφορίες εγγραφών

Σημειώνεται βέβαια σχετικά πως οι εργασίες στο τομέα της εξόρυξης γνώσης στον παγκόσμιο ιστό, μπορούν να χωριστούν σε διάφορες κλάσεις. Η εικόνα που

ακολουθεί δείχνει μια ταξινόμηση των δραστηριοτήτων του τομέα αυτού. Ως ακολούθως, αναλύονται οι τρεις βασικές κατηγορίες για να γίνει περισσότερο κατανοητή η έννοια της εξόρυξης δεδομένων στον παγκόσμιο ιστό.

1.2.1 Μέθοδος Web Content Mining

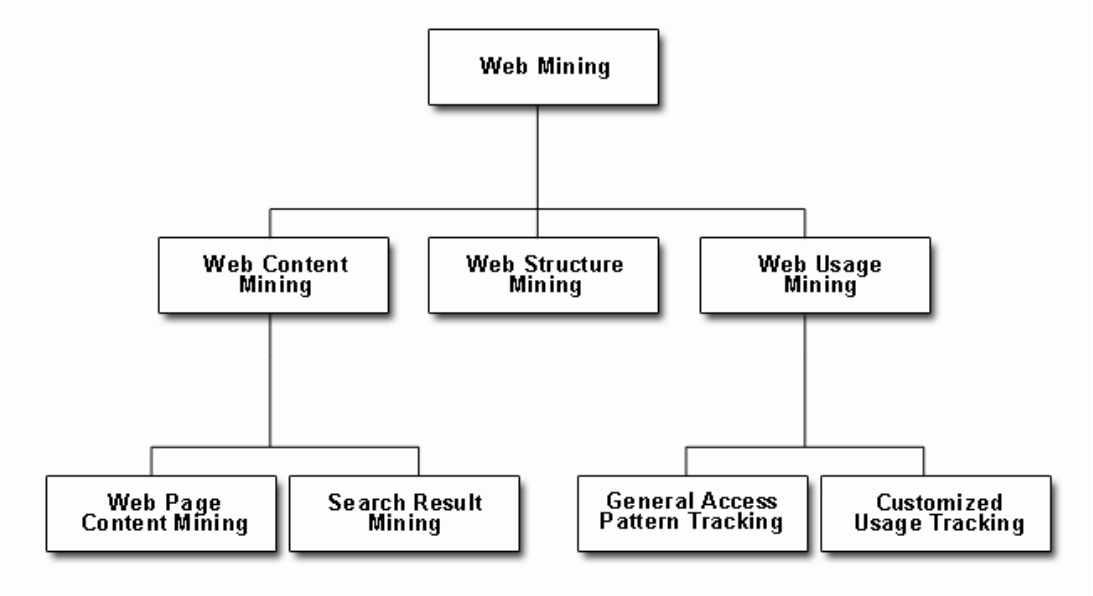
Η μέθοδος του web content mining, εξετάζει τα περιεχόμενα των ιστοσελίδων καθώς και τα αποτελέσματα αναζητήσεων. Το περιεχόμενο αυτό μπορεί να περιέχει τόσο κείμενα όσο και γραφικά. Οι προκλήσεις αυτού του τομέα της εξόρυξης δεδομένων είναι πολλές μιας και το μέγεθος των ιστοσελίδων είναι απροσδιόριστα μεγάλο και η δομή τους δεν είναι ομοιόμορφη. Επίσης υπάρχει πληθώρα κειμένων σε πολλαπλές εκδόσεις καθώς και λανθασμένη και ατελής πληροφορία. Αυτό κάνει ακόμα πιο επιτακτική την ανάγκη για χρήση τεχνικών ώστε τα αποτελέσματα από αναζητήσεις να είναι ορθό και ακριβές (Zanasi, 2003).

Εκτός από αυτό, υπάρχει ένα τμήμα του διαδικτύου γνωστό και ως «βαθύς ιστός (deep web)» το οποίο δεν μπορεί εύκολα να ευρετηριοποιηθεί από μηχανές αναζήτησης. Ο «βαθύς ιστός» περιέχει βάσεις δεδομένων, βιβλιοθήκες, γενετικά δεδομένα και ευρετήρια. Μεγάλο μέρος του «βαθύ ιστού» είναι δομημένο ή ημι-δομημένο και έτσι είναι ευκολότερο να αναλυθεί και να ενοποιηθεί, το δύσκολο είναι να βρεθούν τεχνικές να ευρετηριοποιηθεί.

Η εξόρυξη γνώσης από το περιεχόμενο του Web είναι η διαδικασία της εξαγωγής γνώσης από το περιεχόμενο των Web σελίδων. Ο Cooley και οι άλλοι (71) τη διαχωρίζουν με βάση δύο προσεγγίσεις: την προσέγγιση που βασίζεται σε πράκτορα και τη προσέγγιση που βασίζεται σε βάση δεδομένων. Η πρώτη προσέγγιση, περιλαμβάνει συστήματα τεχνητής νοημοσύνης που μπορούν να ενεργούν αυτόνομα ή ημι-αυτόνομα εκ μέρους ενός συγκεκριμένου χρήστη, για την ανακάλυψη και οργάνωση πληροφορίας που βασίζεται στο Web. Η δεύτερη προσέγγιση επικεντρώνεται στην ομαδοποίηση και οργάνωση ετερογενών και ημι-δομημένων δεδομένων του Web σε περισσότερο δομημένες και υψηλού επιπέδου συλλογές πόρων. Αυτοί οι οργανωμένοι πόροι μπορούν να προσπελαστούν και να αναλυθούν (Berry, Linoff, 2000).

1.2.2 Μέθοδος του Web Structure Mining

Η μέθοδος του web structure mining, είναι ο ερευνητικός τομέας που εστιάζει στη χρήση της ανάλυσης της δομής των συνδέσμων του διαδικτύου, και ένας βασικός σκοπός του είναι η ανακάλυψη των πιο προτιμητέων κειμένων. Ο παγκόσμιος ιστός θεωρείται σαν ένας κατευθυνόμενος γράφος όπου οι ιστοσελίδες είναι οι κόμβοι του και οι σύνδεσμοι είναι οι ακμές που τους ενώνουν. Η βασική ιδέα εδώ είναι πως ένας υπερσύνδεσμος από ένα κείμενο Α σε ένα κείμενο Β υποδηλώνει πως ο συγγραφέας του κειμένου Α θεωρεί το περιεχόμενο του κειμένου Β αξιοσημείωτο. Οι υπερσύνδεσμοι χρησιμοποιούνται ευρέως στις μηχανές αναζήτησης για να αναγνωρίσουν σχέσεις συσχέτισης μεταξύ κειμένων, να ομαδοποιήσουν κείμενα ανάλογα τη σημαντικότητα τους και τελευταία για να βρουν κοινότητες στον παγκόσμιο ιστό από τις παραπομπές ή την μη ύπαρξη παραπομπών (Ramesh et al, 2001).



Εικόνα– Στάδια Εκτέλεσης Διαδικασίας Εξόρυξης Δεδομένων από το Διαδίκτυο– Web Mining. Πηγή: Ramesh et al, 2001

1.2.3 Μέθοδος του Web Usage Mining

Στη μέθοδο του web usage mining γνωστό και ως web log mining, γίνεται επεξεργασία των log αρχείων σχετικά με τις προσβάσεις χρηστών στις διάφορες ιστοσελίδες. Με τη βοήθεια τεχνικών αυτού του τομέα γίνεται κατανοητή η

συμπεριφορά ενός χρήστη αλλά και η δομή της πληροφορίας. Τα δεδομένα των click-stream, τα cookies, τα ερωτήματα των χρηστών, και κάθε είδους δεδομένα σχετικά με τα αποτελέσματα της αλληλεπίδρασης μεταξύ ανθρώπου και διαδικτύου χρησιμοποιούνται επίσης για να τονιστούν οι ανάγκες των πελατών και να βελτιωθεί η ποιότητα των υπηρεσιών τους (Heer, 2004).

Το general access pattern tracking είναι ένας τύπος του web usage mining ο οποίος εξετάζει το ιστορικό επισκέψεων των ιστοσελίδων. Αυτή η χρήση (usage) μπορεί να είναι γενική ή μπορεί να στοχεύει σε συγκεκριμένη χρήση ή χρήστες. Επίσης αναγνωρίζοντας τα πρότυπα της κίνησης, το usage mining και συγκεκριμένα εξόρυξη αυτών ακολουθιακών προτύπων (sequential patterns).

Για παράδειγμα τα πρότυπα μπορούν να συσταδοποιηθούν βάση των ομοιοτήτων τους. Αυτό στη συνέχεια μπορεί να χρησιμοποιηθεί ώστε να γίνει συσταδοποίηση των χρηστών σε ομάδες βασιζόμενοι σε ομοιότητες των προσβάσεων τους σε ιστοσελίδες. Τέλος ένας άλλος τύπος του web usage mining είναι το customized usage tracking το οποίο αναλύει μεμονωμένες τάσεις έτσι ώστε οι ιστοσελίδες να προσδίδονται σε συγκεκριμένους χρήστες. Βασιζόμενοι σε πρότυπα προσβάσεων, μια ιστοσελίδα μπορεί δυναμικά να τροποποιηθεί για ένα χρήστη όσον αφορά την πληροφορία που παρουσιάζει, το βάθος της δομής του και τη μορφή των πηγών που παρουσιάζονται (Banerjee, 2001).

Τελικός σκοπός της παρούσας εργασίας είναι η παρουσίαση ενός μοντέλου για την ομαδοποίηση της περιόδου που εξετάζεται σε υπο-περιόδους ανάλογα με την επισκεψιμότητα της ιστοσελίδας και τα επιμέρους χαρακτηριστικά που περιλαμβάνονται στην έρευνα μας.

1.3 Λόγοι για τους Οποίους Χρησιμοποιείται η Εξόρυξη Δεδομένων σε Διαδικτυακή Βάση

Αναφερόμενοι στους λόγους για τους οποίους χρησιμοποιείται η Εξόρυξη Δεδομένων, στην παρούσα ενότητα, παρουσιάζονται οι βασικές περιοχές εφαρμογής της (Nanopoulos et al, 2011).

1.3.1 Χρήση στο Παγκόσμιο Ιστό

Ο τομέας της εξόρυξης δεδομένων είχε άμεση εφαρμογή με επιτυχία στο Διαδίκτυο. Το πιο δημοφιλές παράδειγμα εξόρυξης δεδομένων στο διαδίκτυο είναι η Google (Berry, Linoff, 2000). Για να γίνει πιο κατανοητή η σημαντικότητα της συνεισφοράς αυτής, θα πρέπει να αντιληφθούμε πως ο όγκος της πληροφορίας που υπάρχει μέχρι τώρα στο διαδίκτυο είναι αδύνατο να μετρηθεί με ακρίβεια. Οι σελίδες που κάθε φορά ερευνά η Google δηλώνεται πως είναι περίπου 4,285,199,774. Κάθε ερώτημα στην μηχανή αναζήτησης δεν ξεπερνά σε χρόνο τα δυο δευτερόλεπτα (Nanopoulos et al, 2011).

Η Google και γενικά ο τομέας της εξόρυξης δεδομένων στο Διαδίκτυο έχουν σήμερα τεράστια επιτυχία γιατί έχουν εκπληρώσει δυο σημαντικούς στόχους. Πρώτα, μπορούν να κάνουν αναζήτηση (με κάθε ερώτημα) σε τόσα πολλά δεδομένα σε πολύ σύντομο χρόνο. Δεύτερον, μπορούν να επιστρέψουν σε κάθε ερώτημα τα πρώτα αποτελέσματα που είναι πιο χρήσιμα. Έτσι τελικά ο χρήστης λαμβάνει γρήγορα και εύκολα μόνο την ουσιώδη πληροφορία που θέλει.

1.3.2 Χρήση σε Διάφορες Οικονομικές και Λοιπές Επιστήμες

Αλγόριθμοι εξόρυξης δεδομένων χρησιμοποιούνται ευρέως σε εφαρμογές από διάφορους άλλους επιστημονικούς τομείς. Ένα αξιοσημείωτο παράδειγμα είναι το SKYCAT (Banerjee, 2001), ένα σύστημα εξόρυξης δεδομένων που αναλαμβάνει ανάλυση και κατηγοριοποίηση χωρικών αντικειμένων. Αυτό που είναι αξιοσημείωτο, είναι πως το SKYCAT εκτελεί αλγόριθμους για την ανίχνευση αντικειμένων από εικόνες.

1.3.3 Χρήση στο Τομέα του Μάρκετινγκ

Μια κατηγορία πολύ γνωστών εφαρμογών εξόρυξης δεδομένων είναι αυτές του μάρκετινγκ. Αυτό είναι αναμενόμενο μιας και μεγάλες εταιρίες χρησιμοποιούν μεγάλα συστήματα διαχείρισης δεδομένων για να διαχειρίζονται μεγάλο αριθμό πελατών και οικονομικών στοιχείων. Τα τελευταία χρόνια οι τάσεις του μάρκετινγκ ορίζουν μια πολιτική έρευνας των αναγκών των πελατών. Αναζητούν απαντήσεις σε ερωτήματα όπως, τι είναι αυτό που θέλουν οι πελάτες, ποιες είναι οι ανάγκες τους

κ.α. Ο τομέας της εξόρυξης δεδομένων έχει συνεισφέρει σημαντικά σε αυτή την κατεύθυνση από την ανάλυση δεδομένων μιας επιχείρησης και την εξαγωγή χρήσιμων συμπερασμάτων για την συμπεριφορά των πελατών (Heer, 2004).

Ένας αρκετά γνωστός αλγόριθμος εξόρυξης δεδομένων είναι ο A-Priori. Ο αλγόριθμος αυτός κάνει ανάλυση δεδομένων αγοράς, όπου υπάρχουν δεδομένα σχετικά με πελάτες ή αγορών σε καταστήματα. Ο A-Priori μπορεί αποδοτικά να δώσει συμπεράσματα όπως «κάθε πελάτης που αγοράζει βαμβακερά υφάσματα θα αγοράσει και μπίρα με μεγάλη πιθανότητα» (Berry, Linoff, 2000). Άλλα παραδείγματα εξόρυξης δεδομένων στο μάρκετινγκ είναι η ανάλυση της συμπεριφοράς των πελατών ηλεκτρονικών καταστημάτων χρησιμοποιώντας τα log αρχεία ή η πρόβλεψη εάν ένας πελάτης θα αγοράσει ένα συγκεκριμένο προϊόν χρησιμοποιώντας παρελθοντικές του κινήσεις.

1.3.4 Χρήση για Πρόβλεψη Επενδύσεων

Πολυάριθμες χρηματιστηριακές εταιρίες χρησιμοποιούν τεχνικές εξόρυξης δεδομένων έτσι ώστε να μπορούν να γνωρίζουν που να επενδύσουν. Στην πραγματικότητα μια μεγάλη μερίδα έρευνας στο τομέα εξόρυξης δεδομένων έχει γίνει έχοντας ως αφετηρία χρηματιστηριακές εφαρμογές. Μια άλλη χρήση των τεχνικών εξόρυξης δεδομένων είναι οι εφαρμογές εξόρυξης δεδομένων από κείμενα. Για παράδειγμα αλγόριθμοι που εξάγουν χρήσιμη πληροφορία από μη δομημένα κείμενα, έτσι ώστε να προβλεφθούν οι τάσεις σε μετοχές (Banerjee, 2001).

1.3.5 Χρήση για Πρόληψη και Ασφάλεια

Η εξόρυξη δεδομένων έχει με επιτυχία εφαρμοστεί και στην πρόληψη και αποφυγή διάφορων τύπων απάτης. Από την αναγνώριση κακόβουλων ενεργειών σε συναλλαγές κάποιος μπορεί να αντιληφθεί συναλλαγές που μπορεί να σχετίζονται με οικονομικές παρανομίες ή άλλου είδους απάτες. Ένα παράδειγμα συστήματος είναι το FAIS (Ramesh et al, 2001).

Ωστόσο τα τελευταία χρόνια, όπως βλέπουμε και ακούμε, υπάρχει μια τάση για πρόληψη σε κακόβουλες ενέργειες. Οι κινήσεις μας σε δημόσιους χώρους

καταγράφεται όπως και αυτές που έχουν να κάνουν με τον παγκόσμιο ιστό. Για παράδειγμα μια πρόσφατη εφαρμογή μπορούσε να αναγνωρίζει ανώμαλα πρότυπα χρησιμοποιώντας κανόνες σε δεδομένα νοσοκομείων έτσι ώστε να αναγνωρίζει, σε πραγματικό χρόνο, εμφάνιση ασθενειών (Banerjee, 2001).

1.4 Στάδια Διαδικασίας

1.4.1 Συλλογή δεδομένων

Τα δεδομένα χρήσης που είναι απαραίτητα για τη διαδικασία μπορούν να συλλεχθούν από διάφορες πηγές. Οι εφαρμογές στον τομέα της μεταλλευτικής χρήσης του web βασίζονται σε δεδομένα που συλλέγονται από 3 βασικές πηγές:

1) Διακομιστές διαδικτύου

Οι διακομιστές διαδικτύου είναι η πιο μεγάλη πηγή δεδομένων. Μπορούν να συλλέξουν μεγάλους όγκους δεδομένων στα log αρχεία τους και στις βάσεις δεδομένων τους. Αυτά τα αρχεία συνήθως περιέχουν βασικές πληροφορίες όπως το όνομα, την IP διεύθυνση, ημερομηνία και ο χρόνος που έγινε το αίτημα. Αυτή η πληροφορία συνήθως αναπαρίσταται σε κάποια προκαθορισμένη μορφή: κοινή μορφή αρχείου καταγραφής, εκτεταμένη μορφή αρχείου καταγραφής, logml. Σε μερικές περιπτώσεις χρησιμοποιούνται βάσεις δεδομένων αντί για αρχεία για την αποθήκευση όλων αυτών των πληροφοριών έτσι ώστε να μπορούν να εφαρμοστούν πιο αποδοτικά ερωτήματα πάνω στα δεδομένα.

Το σημαντικότερο ζήτημα κατά την χρήση των δεδομένων από τα log αρχεία των διακομιστών είναι η αναγνώριση των συνόδων των χρηστών (π.χ. πως να ομαδοποιήσεις όλα τα αιτήματα για ιστοσελίδες (ή των κλικ ρευμάτων) από τους χρήστες για την εύρεση των μονοπατιών που διέσχισε ο χρήστης κατά την περιήγησή του σε μια ιστοσελίδα). Αυτή η εργασία είναι πολλές φορές δύσκολη και δαπανηρή τόσο σε χρόνο όσο και σε χώρο και εξαρτάται πολύ από τον τύπο της πληροφορίας που ένας διακομιστής μπορεί να φυλάξει. Η πιο κοινά χρησιμοποιούμενη τεχνική είναι η χρήση των cookies για την καταγραφή των ακολουθιών των αιτημάτων για ιστοσελίδες. Εάν τα cookies δεν είναι διαθέσιμα, υπάρχουν πολλές ευρετικές μέθοδοι (μέθοδοι heuristics) που μπορούν να εφαρμοστούν για την εύρεση των συνόδων κάθε χρήστη. Ωστόσο ακόμα και αν χρησιμοποιηθούν τα cookies, είναι μερικές φορές αδύνατο να γίνουν γνωστές οι ακριβείς κινήσεις των χρηστών σε μια ιστοσελίδα μιας και η προς τα πίσω κίνηση ενός χρήστη (πίσω) δεν καταγράφεται στο διακομιστή.

Εκτός από τα log αρχεία, η συμπεριφορά του χρήστη μπορεί να καταγραφεί επίσης από την πλευρά του διακομιστή από τα πακέτα TCP / IP. Ακόμα και σε αυτή την περίπτωση η αναγνώριση των χρηστών είναι ένα θέμα, αλλά η χρήση των πακέτων παρέχει πολλά πλεονεκτήματα. Συγκεκριμένα 1) τα δεδομένα συλλέγονται σε πραγματικό χρόνο, 2) πληροφορίες που προέρχονται από διαφορετικούς διακομιστές μπορούν να ενωθούν σε κοινά αρχεία και 3) η χρήση ειδικών κουμπιών όπως αυτού της προς τα πίσω περιήγησης (πίσω) μπορούν να ανιχνευθούν. Παρά τα πλεονεκτήματα η μέθοδος αυτή χρησιμοποιείται πολύ σπάνια στην πράξη γιατί σε διακομιστές με μεγάλη κίνηση παρουσιάζονται προβλήματα κλιμάκωσης και πληροφορίες που είναι κρυπτογραφημένες δεν είναι δυνατόν να αναγνωριστούν. Η πιο καλή μέθοδος για την καταγραφή της χρησιμοποίησης διαδικτύου είναι η απευθείας πρόσβαση στα δεδομένα του διακομιστή. Δυστυχώς αυτό δεν είναι πάντα εφικτό.

2) Διακομιστές μεσολάβησης

Πολλοί πάροχοι υπηρεσιών διαδικτύου (ISPs) δίνουν στους πελάτες τους δυνατότητες χρήσης υπηρεσιών μεσολάβησης διακομιστών μεσολάβησης (proxy servers για) για τη βελτίωση της ταχύτητας περιήγησης με τη χρήση κρυφής μνήμης. Γενικά η συλλογή δεδομένων περιήγησης στον διακομιστή είναι ίδια με αυτή που γίνεται σε οποιοδήποτε διακομιστή διαδικτύου. Η μόνη διαφορά σε αυτή την περίπτωση είναι πως στο proxy διακομιστή συλλέγονται δεδομένα από ομάδες χρηστών που έχουν πρόσβαση σε τεράστιες ομάδες από διακομιστές διαδικτύου. Ακόμα και σε αυτή την περίπτωση, η ανακατασκευή των συνόδων (συνεδρίες) είναι δύσκολη και δεν είναι δυνατή η ανίχνευση όλων των μονοπατιών που έχουν κάνει οι χρήστες. Ωστόσο, όταν δεν υπάρχει άλλη μέθοδος χρήσης προσωρινής μνήμης μεταξύ των χρηστών και proxy διακομιστή, η αναγνώριση των συνόδων των χρηστών είναι πιο εύκολη.

3) Χρήστες διαδικτύου

Τα δεδομένα χρήσης μπορούν να καταγραφούν επίσης και από την πλευρά του χρήστη χρησιμοποιώντας Javascript γλώσσα ή βοηθητικές εφαρμογές Java. Αυτές οι τεχνικές αποφεύγουν τα προβλήματα αναγνώρισης των συνόδων των χρηστών και τα προβλήματα που προκαλούνται από τη χρήση κρυφής μνήμης (όπως η χρήση του κουμπιού «πίσω»). Επίσης, παρέχουν αναλυτικές πληροφορίες για την συμπεριφορά των χρηστών. Ωστόσο, αυτές οι προσεγγίσεις βασίζονται πολύ στη συνεργασία του χρήστη και προκύπτουν πολλά ζητήματα σχετικά με νόμους ιδιωτικότητας.

1.4.2 Προεπεξεργασία δεδομένων

Η προεπεξεργασία των δεδομένων παίζει σπουδαίο ρόλο στον τομέα της μεταλλευτικής χρήσης του web και τις εφαρμογές του. Ειδικά η προεπεξεργασία των web log δεδομένων είναι συνήθως πολύπλοκη και απαιτεί πολύ χρόνο. Αποτελείται από τέσσερα διαφορετικά βήματα:

1) Καθαρισμός δεδομένων: Αυτό το βήμα αποτελείται από την διαγραφή όλων των δεδομένων από τα web log που δεν είναι χρήσιμα για τους σκοπούς της ανάλυσης. Για παράδειγμα τα αιτήματα για αρχεία εικόνας και άλλων αρχείων που περιέχονται σε σελίδες αλλά και συνόδων περιήγησης που προέρχονται από αυτόματους μηχανισμούς (ρομπότ) και web crawlers. Ενώ τα αρχεία που περιέχονται σε σελίδες είναι εύκολο να αφαιρεθούν οι αυτόματοι μηχανισμοί και οι αράχνες Ιστού θα πρέπει ρητώς να αναγνωριστούν.

2) Αναγνώριση και ανακατασκευή των συνόδων των χρηστών: Αυτό το βήμα αποτελείται από την αναγνώριση των συνόδων των διαφορετικών χρηστών από τη συνήθως μικρή πληροφορία που παρέχουν τα web log αρχεία και από την ανακατασκευή των μονοπατιών των χρηστών μέσα από τις αναγνωρισμένες συνόδους. Η πολυπλοκότητα του βήματος αυτού δεν είναι συγκεκριμένη και εξαρτάται από την ποιότητα της παρεχόμενης πληροφορίας από τα web log αρχεία. Τα περισσότερα προβλήματα εδώ προέρχονται από την επαναποθήκευση (caching) που γίνεται είτε στους proxy διακομιστές είτε στα προγράμματα περιήγησης (browsers). Για παράδειγμα η επαναποθήκευση που γίνεται στους διακομιστές μεσολάβησης (proxy servers) που μπορεί να προκαλέσει μια IP διεύθυνση να σχετίζεται με διαφορετικές συνόδους του χρήστη και έτσι γίνεται αδύνατη η χρήση της IP διεύθυνσης σαν αναγνωριστικό του χρήστη.

Αυτό το πρόβλημα μπορεί να αντιμετωπιστεί σε κάποιο βαθμό με τη χρήση αρχείων cookies, επανεγγραφή της ηλεκτρονικής διεύθυνσης, ή απαιτώντας από το χρήστη να κάνει σύνδεση στις σελίδες με κάποιο συνθηματικό. Το αρχείο cookie είναι ένα αρχείο με δεδομένα που στέλνεται από το διακομιστή διαδικτύου στο πρόγραμμα περιήγησης. Αυτή η πληροφορία αποθηκεύεται στον υπολογιστή του χρήστη σαν αρχείο κειμένου. Τα αρχεία αυτά μπορούν να περιέχουν πολλές πληροφορίες για τους χρήστες και μεταξύ αυτών το αναγνωριστικό του χρήστη το οποίο χρειαζόμαστε. Αυτή η πληροφορία μπορεί να παρέχεται στο διακομιστή κάθε φορά που αυτός το ζητάει και να αποθηκεύεται στο web log αρχείο μαζί με το αίτημα της σελίδας.

Υπάρχουν ωστόσο περιπτώσεις που η τεχνική αυτή δε λειτουργεί. Μερικά προγράμματα περιήγησης δεν υποστηρίζουν τη χρήση των cookies ή έχουν ειδικές ρυθμίσεις που μπορεί κάποιος να τα έχει απενεργοποιήσει. Σε αυτές τις περιπτώσεις, η επανεγγραφή της ηλεκτρονικής διεύθυνσης μπορεί να χρησιμοποιηθεί για την εγγραφή του κωδικού της συνόδου στο τέλος της ηλεκτρονικής διεύθυνσης.

3) Ανάκτηση δεδομένων σχετικά με το περιεχόμενο και τη δομή των σελίδων: Η πλειοψηφία των εφαρμογών web εξόρυξης χρησιμοποιούν τις διευθύνσεις για τις οποίες έχουν γίνει αιτήματα σαν τη βασική πηγή δεδομένων για να εφαρμόσουν τεχνικές εξόρυξης δεδομένων. Οι διευθύνσεις ωστόσο δεν προσφέρουν πολύ πληροφορία μιας και δε μεταφέρουν πληροφορίες σχετικά με το περιεχόμενο των σελίδων.

4) Μορφοποίηση των δεδομένων: Αυτό είναι το τελικό στάδιο της προεπεξεργασίας. Αφού έχουν ολοκληρωθεί οι προηγούμενες φάσεις, τα δεδομένα μορφοποιούνται πριν εφαρμοστεί σε αυτά κάποια τεχνική εξόρυξης γνώσης.

1.4.3 Ανακάλυψη προτύπων

Αυτό είναι το σημαντικότερο στάδιο της διαδικασίας, καθώς εδώ γίνεται η ανακάλυψη της επιθυμητής γνώσης από τα δεδομένα. Για το σκοπό αυτό, χρησιμοποιούνται τεχνικές από τη μηχανική μάθησης και τη στατιστική. Για τις εφαρμογές εξατομίκευσης στον παγκόσμιο ιστό, ως γνώση θεωρούνται κάποια πρότυπα που αντικατοπτρίζουν τη συμπεριφορά των χρηστών ως προς την περιήγησή τους στον ιστό.

Οι πιο πολλές από τις εμπορικές εφαρμογές στον τομέα του μεταλλευτική χρήση του web εκμεταλλεύονται συνδυασμούς τεχνικών ανάλυσης. Αντίθετα η έρευνα σε αυτή την περιοχή περισσότερο εστιάζει στην δημιουργία τεχνικών ανακάλυψης γνώσης για την ανάλυση των δεδομένων web χρήση. Οι περισσότερες από αυτές τις προσπάθειες επικεντρώνονται σε τρία παραδείγματα:

1) Κανόνες αυτοσυσχέτισης: Αυτή είναι ίσως η πιο στοιχειώδης τεχνικής εξόρυξης δεδομένων και την ίδια στιγμή η πιο συχνά χρησιμοποιούμενη τεχνική στον τομέα της web εξόρυξης. Κατά την εφαρμογή κανόνων αυτοσυσχέτισης στον τομέα της web εξόρυξης βρίσκουμε συσχετίσεις μεταξύ των σελίδων που εμφανίζονται μαζί στις συνόδους των χρηστών. Ένα τυπικό παράδειγμα έχει τη μορφή: "A.html, B.html → C.html". Εδώ δηλώνεται πως εάν ένας χρήστης έχει επισκεφτεί τη σελίδα A.html και τη σελίδα B.html είναι πολύ πιθανό να επισκεφτεί και την σελίδα C.html.

2) Πρότυπα ακολουθιών: Τα πρότυπα ακολουθιών χρησιμοποιούνται για την εύρεση συχνών τμημάτων ακολουθιών μεταξύ μεγάλων όγκων ακολουθιών δεδομένων. Στον τομέα του web εξόρυξη χρήση, χρησιμοποιούνται για να βρούμε πρότυπα ακολουθιών περιήγησης που παρουσιάζονται συχνά στις συνόδους των χρηστών.

3) Συσταδοποίηση: Είναι μια τεχνική που βρίσκει ομάδες όμοιων αντικειμένων μεταξύ μεγάλων όγκων δεδομένων χρησιμοποιώντας την ιδέα της απόστασης η οποία δείχνει την ομοιότητα μεταξύ αντικειμένων. Η τεχνική αυτή χρησιμοποιείται πολύ στον τομέα του μεταλλευτική χρήση του web για την ομαδοποίηση πολλών όμοιων συνόδων. Στη συσταδοποίηση έχει προταθεί να μην γίνεται εστίαση στις συνόδους συγκεκριμένων χρηστών αλλά στην ομαδοποίηση συνόδων από διαφορετικούς χρήστες. Επίσης έχει προταθεί μια τεχνική που ψάχνει για ομοιότητες μεταξύ γραφημάτων σε συνδυασμό με το χρόνο που έχει ξοδευτεί σε κάθε σελίδα.

1.4.4 Εκμετάλλευση της γνώσης

Στο τελευταίο αυτό στάδιο γίνεται η ερμηνεία και η αξιολόγηση της γνώσης που έχει εξαχθεί και παρουσιάζεται σε κατανοητή μορφή. Επίσης, η γνώση αυτή αξιοποιείται σε συστήματα εξατομίκευσης στον ιστό.

1.5 Οφέλη που Προκύπτουν Αναφορικά με την Ανάλυση Στοιχείων για Χρήστες του Διαδικτύου και το Ενδιαφέρον τους για τα Προϊόντα – Υπηρεσίες των Διαφημιζόμενων Επιχειρήσεων

Αναφερόμενοι σχετικά στα οφέλη που προκύπτουν για τους χρήστες από τη μέθοδο εξόρυξης δεδομένων, θα λέγαμε πως για έναν ιδιοκτήτη ή διαχειριστή κάποιου ιστότοπου, υπάρχουν κάποιες μεταβλητές, οι οποίες είναι πολύ σημαντικές και οι οποίες περιγράφουν τη λειτουργία και τη διαχείριση τόσο του ιστότοπου γενικότερα, όσο και της κάθε σελίδας αυτού ειδικότερα. Συνεπώς είναι πιο επίκαιρο από ποτέ το πώς θα μπορέσουν να προβλεφθούν οι μελλοντικές τιμές για τις μεταβλητές αυτές οι οποίες κατά πολύ μπορούν να δείξουν την επιτυχία του κάθε site τη δεδομένη στιγμή και προβλέψεις για το μέλλον (Zanasi, 2003).

Τέτοιου είδους μεταβλητές επί παραδείγματι, μπορούν να είναι οι ακόλουθες (Nanopoulos et al, 2011):

- Ο Αριθμός των επισκεπτών στο site
- Ο συνολικός χρόνος παραμονής στο site αλλά και στη κάθε σελίδα αυτού
- Κατά πόσον ο κάθε επισκέπτης φτάνει στους στόχους του κάθε site

Για τη συλλογή των δεδομένων που θα χρειαστούν για την έρευνα ενός ατόμου σχετικά με ένα τομέα εμπορικής λειτουργίας, μπορούν να χρησιμοποιηθούν δεδομένα τα οποία προκύπτουν από εργαλεία Ανάλυσης Δεδομένων του Ιστοτόπου. Το πλέον διαδεδομένο από αυτά τα εργαλεία είναι το Google Analytics. Το Google Analytics στην πράξη είναι μία δωρεάν εφαρμογή που παρέχει αναλυτικά στοιχεία σχετικά με την επισκεψιμότητα ενός ιστότοπου. Χωρίς στατιστικά στοιχεία δεν μπορούμε να γνωρίζουμε πόσοι άνθρωποι επισκέπτονται το site, τι επισκέπτονται περισσότερο / λιγότερο, από ποιες χώρες προέρχονται οι επισκέπτες, πόσος είναι ο μέσος χρόνος επίσκεψης κ.ο.κ. (Banerjee, 2001)

Κάποια χρήσιμα στοιχεία που προσφέρονται από τα Google Analytics μέσω της χρήσης εξόρυξης δεδομένων, είναι τα ακόλουθα (Berry, Linoff, 2000):

- *Site Usage*: Μέσω αυτού βλέπουμε το διάστημα ημερομηνιών. Το διάστημα αυτό μπορεί να οριστεί σύμφωνα με τις ανάγκες μας. Ορίζοντας το χρονικό πλαίσιο μπορούμε να δούμε τα στατιστικά μόνο για τη συγκεκριμένη χρονική περίοδο που ορίσαμε
- *Visits/Visitors*: Δείχνει τον αριθμό επισκέψεων. Ο αριθμός αυτός δεν ορίζει τους μοναδικούς επισκέπτες. Για να γίνει καλύτερα αντιληπτό αυτό αναφέρεται το ακόλουθο παράδειγμα: «Εάν ένας επισκέπτης έρθει στο site 3 φορές στο χρονικό διάστημα που αναλύει η αναφορά τότε καταμετρείται 3 φορές»
- *Unique Visits/Visitors*: Δείχνει τον αριθμό των μοναδικών επισκέψεων. Οι μοναδικοί επισκέπτες μετρώνται βάση της μοναδικής IP ή του cookie, η οποία είναι η ηλεκτρονική ταυτότητα του κάθε επισκέπτη
- Pageviews: Συνολικός αριθμός σελίδων που επισκέφτηκαν οι επισκέπτες που ορίζονται από την μεταβλητή Visits, όπως αυτή αναφέρθηκε παραπάνω.
- *Pages/Visit*: Μέσος όρος σελίδων που επισκέπτεται ένας επισκέπτης.

- Bounce Rate: Ποσοστό αναπήδησης που ορίζει το ποσοστό των χρηστών που βλέπουν μόνο μια σελίδα και στη συνέχεια εγκαταλείπουν (αναπηδούν) το site, χωρίς να έχουν την οποιαδήποτε αλληλεπίδραση με τις λειτουργίες της σελίδας. Στην ουσία είναι από τους πλέον σημαντικούς δείκτες για την επιτυχία ενός site και επηρεάζεται από το πόσο ενδιαφέρον βρίσκει ο χρήστης το περιεχόμενο της σελίδας που επισκέπτεται.
- *Avg. Time on Site*: Μέσος χρόνος επίσκεψης.
- *New Visits*: Ποσοστό νέων επισκέψεων.
- *Returning Visits*: Ποσοστό επισκεπτών που έχουν και κατά το παρελθόν επισκεφτεί τον ιστότοπο.
- *Map Overlay*: Μας δείχνει τη γεωγραφική κατανομή των επισκεπτών του ιστότοπου. Η ανάλυση μπορεί να γίνει τόσο σε επίπεδο χώρας, όσο και σε επίπεδο πόλης. Προσαρμόζοντας τα κατάλληλα φίλτρα, μας δίνεται η δυνατότητα μεγαλύτερης πληροφόρησης σχετικά με τον αριθμό των επισκεπτών από την κάθε πόλη, του μέσου χρόνου επίσκεψης κτλ.
- *Traffic Sources Overview*: Στο πεδίο αυτό εμφανίζονται οι πηγές από τις οποίες έρχονται στο συγκεκριμένο ιστότοπο οι επισκέπτες. Τέτοιες πηγές μπορούν να είναι είτε οι μηχανές αναζήτησης (Search Engines), είτε η απευθείας πληκτρολόγηση του URL (Uniform Resource Locator: δηλώνει μια διεύθυνση ενός πόρου του Παγκόσμιου Ιστού), είτε παραπομπές από άλλους ιστότοπους (Referral Links), είτε τα Μέσα Κοινωνικής Δικτύωσης (Social Media)
- *Conversions*: Η αυτολεξής μετάφραση είναι Μετατροπές. Στην πράξη ο όρος αυτός δηλώνει τον αριθμό των στόχων που έχει κάποιος ιστότοπος. Οι στόχοι αυτοί διαφοροποιούνται από ιστότοτοπο σε ιστότοπο. Επί παραδείγματι, δύναται να είναι η πώληση ενός αγαθού ή μίας υπηρεσίας σε ιστότοπους ηλεκτρονικού εμπορίου (e-commerce) ή η συμπλήρωση μίας φόρμας ή το download (διαδικασία απόκτησης μέσω μεταφοράς δεδομένων από κάποιον ιστότοπο) κάποιου αρχείου
- *Conversion Funnel*: (Μονοπάτι Διοχεύτεσης Μετατροπής) Τα βήματα που ακολουθούνται από τον επισκέπτη ενός site έως ότου ολοκληρωθεί μια μετατροπή π.χ αγορά προϊόντος, εγγραφή σε newsletter κλπ

- *Conversion Rate*: (Ποσοστό Μετατροπής) Το ποσοστό των επισκεπτών που εκτελούν μια μετατροπή και είναι το αποτέλεσμα της διαίρεσης με το συνολικό αριθμό επισκεπτών του ιστότοπου.

Είδαμε παραπάνω ότι από το πεδίο Traffic Sources Overview μπορούμε να δούμε την πηγή από την οποία ήλθε στον ιστότοπο ο κάθε επισκέπτης. Στην ουσία το πεδίο αυτό μας δείχνει τα αποτελέσματα της διαφήμισης του ιστότοπου στο διαδίκτυο. Η διαφήμιση ενός ιστότοπου στοχεύει σε 2 φάσεις. Η 1η φάση στοχεύει στο να μετατρέψει τον αποδέκτη σε επισκέπτη, και η 2η φάση να μετατρέψει τον επισκέπτη σε πελάτη (η έννοια πελάτης ποικίλει, αναλόγως με τον τύπο του Conversion που έχει δηλωθεί).

Η διαφήμιση στο διαδίκτυο μπορεί να έχει διάφορες μορφές αναλόγως αφενός του Budget του διαφημιζόμενου και αφετέρου του σκοπού που θέλει να επιτευχθεί ο διαφημιζόμενος (Awareness VS Performance δηλαδή Αναγωρισιμότητα εναντίον Πωλήσεων).

Ωστόσο ο τρόπος εκείνος διαφήμισης ο οποίος παρουσιάζει το μεγαλύτερο στατιστικά ενδιαφέρον είναι οι SEM καμπάνιες[1]. Οι καμπάνιες αυτές χωρίζονται σε 2 κατηγορίες (Ramesh et al, 2001):

- Στο Search, όπου εμφανίζονται διαφημίσεις αναλόγως του ερωτήματος (Query) που έχει γίνει στη μηχανή αναζήτησης και η στόχευση γίνεται βάσει Keywords (λέξεις κλειδιά)
- Στο GDN (Google Display Network, το οποίο είναι ένα παγκόσμιο δίκτυο συνεργαζομένων με τη Google ιστότοπων), όπου η λειτουργία του μοιάζει αρκετά με αυτή του Premium Display.

Τέλος, σημειώνεται πως το περιβάλλον για τις SEM καμπάνιες είναι το Google Adwords που είναι η πλατφόρμα διαχείρισης τους. Το βασικό τους πλεονέκτημα είναι ότι εν αντιθέσει του Premium Display, προσεγγίζονται άτομα τα οποία αναζητούν αυτό το οποίο προσφέρει ο διαφημιζόμενος ιστότοπος, καθώς και

[1] Το Search Engine Marketing ή αλλιώς το Μάρκετινγκ των μηχανών αναζήτησης (SEM), είναι μια μορφή του μάρκετινγκ στο διαδίκτυο που αποβλέπει στην προώθηση των ιστοσελίδων και την αύξηση της προβολής τους στα αποτελέσματα των μηχανών αναζήτησης (SERPs). Αυτό επιτυγχάνεται συνήθως είτε μέσω της χρήσης πληρωμένης κατάταξης στα αποτελέσματα, είτε μέσω θεματικής διαφήμισης

ότι η χρέωση υφίσταται από τη στιγμή που κάποιος χρήστης αλληλεπιδράσει με τη διαφήμιση και όχι μόνο με την εμφάνισή της.

1.6 Παράγοντες του Διαδικτύου που Επηρεάζουν την Εξόρυξη Δεδομένων

Υπάρχουν δύο παράγοντες επιτυχίας στην εξόρυξη δεδομένων. Ο πρώτος είναι η ακριβής διατύπωση του προβλήματος. Μια επικεντρωμένη πρόταση έχει συνήθως τα καλύτερα αποτελέσματα. Ο δεύτερος παράγοντας-κλειδί είναι τα ορθά δεδομένα (Zanasi, 2003). Η εξόρυξη δεδομένων δεν παρέχει αυτομάτως λύσεις χωρίς καθοδήγηση. Εκτός αυτού, αν και ένα καλό εργαλείο εξόρυξης δεδομένων προστατεύει από περίπλοκες στατιστικές τεχνικές, είναι απαραίτητη η κατανόηση του τρόπου λειτουργίας των εργαλείων που επιλέγονται και των αλγορίθμων επί των οποίων βασίζονται. Όπως και με όλες τις τεχνικές διαχείρισης γνώσης, η χρήση τόσο των ορθών δεδομένων (ρητή γνώση) όσο και η καλή τεχνογνωσία που αφορά την επιχειρηματική λειτουργία (άρρητη γνώση) έχουν μεγάλη σημασία (Banerjee, 2001).

Μεγάλος αριθμός εταιρειών έχουν αναπτύξει επιτυχείς εφαρμογές εξόρυξης δεδομένων. Ενώ οι πρώτοι που υιοθέτησαν την τεχνολογία αυτή ανήκαν κυρίως στον τομέα έντασης πληροφοριών, όπως οι χρηματο-οικονομικές υπηρεσίες και το μάρκετινγκ άμεσης ταχυδρόμησης, η τεχνολογία είναι εφαρμόσιμη σε οποιαδήποτε εταιρεία αναζητά να χρησιμοποιήσει αποδοτικά μια μεγάλη αποθήκη δεδομένων ώστε να διαχειριστεί με καλύτερο τρόπο τις πελατειακές της σχέσεις.

Το μέγιστο επιχειρηματικό όφελος επιτυγχάνεται από την διάθεση των μοντέλων που προκύπτουν μέσω της διαδικασίας της εξόρυξης των δεδομένων (Data Mining) στα σημεία επαφής της επιχείρησης με τον πελάτη (διαδίκτυο, καταστήματα πώλησης, τηλεφωνικό κέντρο, γραπτή επικοινωνία κ.λ.π), οπότε μπορούμε να απαντάμε ερωτήματα της μορφής "τι θα μπορούσαμε να προσφέρουμε στον συγκεκριμένο πελάτη σήμερα για να τον διατηρήσουμε ενεργό στην επιχείρηση" (Heer, 2004).

Αρκετό ενδιαφέρον παρουσιάζουν επίσης και τα εργαλεία εξόρυξης δεδομένων τα οποία εξελίσσονται συνεχώς βασισμένα σε ιδέες από τις τελευταίες επιστημονικές έρευνες. Πολλά από τα εργαλεία αυτά ενσωματώνουν τους πιο

πρόσφατους αλγορίθμους από την τεχνητή νοημοσύνη, τη στατιστική και τη βελτιστοποίηση. Προς το παρόν η γρήγορη επεξεργασία επιτυγχάνεται με χρήση σύγχρονων τεχνικών βάσεων δεδομένων ,όπως κατανεμημένη επεξεργασία με αρχιτεκτονικές client/server με παράλληλες βάσεις δεδομένων και με αποθήκες δεδομένων. Η μελλοντική τάση είναι προς την ανάπτυξη πιο πλήρων διαδικασιών διαδικτύου. Η επεξεργασία πρέπει να εκτελείται με χρήση όλων των διαθέσιμων πηγών. Η ύπαρξη κατανεμημένων περιβαλλόντων, παρέχοντας τη δυνατότητα κατανομής των πόρων όλων των συστημάτων θα ωφελήσει τη διαδικασία της επεξεργασίας των δεδομένων ως προς το χρόνο και τη μνήμη (Ramesh et al, 2001).

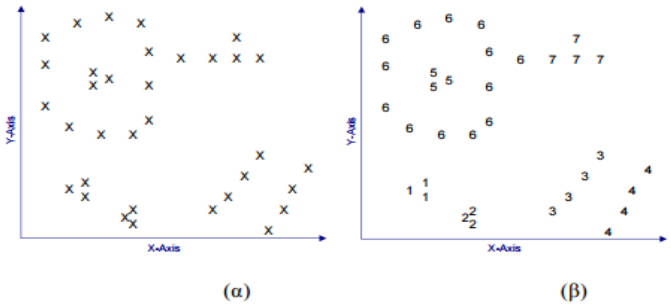
ΚΕΦΑΛΑΙΟ 2

Τεχνικές ομαδοποίησης σε διαδικτυακά δεδομένα

2.1 Εισαγωγή

Η ανάλυση των δεδομένων αποτελεί τη βάση σε πολλές εφαρμογές στον τομέα της πληροφορικής, είτε κατά την διάρκεια της σχεδίασης κάποιας εφαρμογής ή κατά την λειτουργία της. Οι διαδικασίες ανάλυσης δεδομένων μπορούν να χωριστούν σε δύο κατηγορίες, τις διαδικασίες ανεύρεσης, με στόχο την ανακάλυψη και την κατασκευή υποθέσεων (hypothesis) από τα δεδομένα, και τις διαδικασίες επιβεβαίωσης με στόχο την λήψη αποφάσεων δεδομένης της δομής της πληροφορίας.

Ο διαχωρισμός βασίζεται στην ύπαρξη ή μη κατάλληλων μοντέλων τα οποία εκφράζουν την πηγή των δεδομένων. Και στις δύο περιπτώσεις διαδικασιών όμως, σημείο κλειδί είναι η ομαδοποίηση (clustering) των στοιχείων με βάση (α) το υιοθετημένο μοντέλο, ή (β) τις φυσικές ομάδες δεδομένων που προκύπτουν από την ανάλυση των δεδομένων. Η ανάλυση συστάδων (cluster analysis) ή πιο απλά το clustering είναι η οργάνωση μιας συλλογής από δείγματα-στοιχεία (patterns) σε συστάδες (clusters) με βάση κάποιο μέτρο ομοιότητας. Τα στοιχεία συνήθως περιγράφονται σαν διανύσματα τιμών κάποιων μέτρων ή αναπαριστώνται ως σημεία σε έναν πολυδιάστατο χώρο. Στοιχεία που ανήκουν στην ίδια ομάδα παρουσιάζουν μεγαλύτερη ομοιότητα από ότι στοιχεία που ανήκουν σε διαφορετικές ομάδες. Ένα παράδειγμα ομαδοποίησης φαίνεται στο σχήμα παρακάτω. Τα δεδομένα μας είναι αυτά που φαίνονται στο σχήμα (α) και η ομαδοποίηση τους φαίνεται στο σχήμα (β). Στοιχεία που ανήκουν στην ίδια ομάδα φέρουν κοινό αριθμό. Η ποικιλία τεχνικών για την αναπαράσταση των δεδομένων, έκφρασης της ομοιότητας μεταξύ στοιχείων και ομαδοποίησης των δεδομένων έχει ως αποτέλεσμα την ύπαρξη μιας πλούσιας συλλογής μεθόδων ομαδοποίησης.



Ομαδοποίηση δεδομένων (data clustering. Πηγή: Nigam et al., 2000

Το Clustering είναι πολύ χρήσιμο σε πολλούς τομείς όπως η ανάλυση προτύπων (pattern-analysis), η λήψη αποφάσεων (decision-making), η μηχανική εκμάθηση (machine-learning), η εξόρυξη δεδομένων (data mining), η ανάκτηση κειμένων (document retrieval) κ.α. Στις περισσότερες των περιπτώσεων που εφαρμόζεται το Clustering υπάρχει μικρή γνώση για την δομή και το είδος των στοιχείων π.χ. στατιστικά μοντέλα, που να περιγράφουν τα δεδομένα. Έτσι ο υπεύθυνος για την λήψη των τελικών αποφάσεων και την εφαρμογή του Clustering στα δεδομένα θα πρέπει να κάνει κάποιες υποθέσεις για τα δεδομένα. Κάτω από αυτούς τους περιορισμούς η μεθοδολογία του Clustering διαφαίνεται ιδιαίτερα κατάλληλη για την ανακάλυψη αλληλοσυσχετισμών μεταξύ των δεδομένων προκειμένου να κατανοηθεί η δομή τους, κάτι που είναι και ο απώτερος στόχος.

2.2 Η διαδικασία του Clustering

Η συσταδοποίηση είναι η προσπάθεια της ομαδοποίησης ενός συνόλου αντικειμένων με τέτοιο τρόπο ώστε τα αντικείμενα της ίδιας ομάδας (που ονομάζεται συστάδα – cluster) είναι ομοιότερα (κατά κάποια έννοια) από αυτά των υπολοίπων ομάδων. Είναι μια κύρια εργασία της εξόρυξης γνώσης και μια συνήθης τεχνική της στατιστικής ανάλυσης, χρησιμοποιείται δε σε πολλά πεδία επιστημών συμπεριλαμβανομένων της μηχανικής μάθησης, της αναγνώρισης προτύπων, της ανάλυσης εικόνων, της ανάκτησης πληροφοριών και της βιοτεχνολογίας. Υπάρχουν πολλοί αλγόριθμοι συσταδοποίησης που διαφέρουν όμως σημαντικά στην έννοια του τι αποτελεί μια συστάδα και ποιος είναι ο πιο αποδοτικός τρόπος εύρεσής τους.

Δημοφιλείς αντιλήψεις των συστάδων περιλαμβάνουν ομάδες με μικρές αποστάσεις μεταξύ των μελών τους, πυκνές περιοχές στο χώρο των δεδομένων, διαστήματα ή συγκεκριμένες στατιστικές κατανομές. Παρόλο λοιπόν που στη

μηχανική μάθηση κατατάσσεται στις μεθόδους μη εποπτευόμενης μάθησης, η αλήθεια είναι ότι υπάρχουν παράμετροι σε κάθε αλγόριθμο που αν ρυθμιστούν σωστά μπορεί να φέρουν καλύτερα χαρακτηριστικά στο συγκεκριμένο σύνολο δεδομένων. Επίσης είναι δυνατό η συσταδοποίηση να χρησιμοποιηθεί και για κατηγοριοποίηση, όπου μετά την εύρεση των συστάδων από τον αλγόριθμο, η περαιτέρω τροφοδοσία με δεδομένα μας προβλέπει σε ποια συστάδα θα ανήκουν αυτά.

Η συσταδοποίηση χρησιμοποιήθηκε αρχικά στην ανθρωπολογία από τους Driver και Kroeber το 1932 και αργότερα στην ψυχολογία από τον Zubin το 1938. Όπως είναι φανερό η έννοια της συστάδας είναι δύσκολο να οριστεί με ακρίβεια πέραν του ότι είναι μια ομάδα αντικειμένων. Έτσι λοιπόν ανάλογα με την έννοια που δίνει κάποιος στη συστάδα προκύπτουν και οι διαφορετικοί αλγόριθμοι της συσταδοποίησης. Τυπικά μοντέλα συσταδοποίησης είναι τα:

• Μοντέλα συνδεσιμότητας (Connectivity models). Περιλαμβάνουν την ιεραρχική συσταδοποίηση που δημιουργεί μοντέλα ανάλογα με το μήκος της διασύνδεσης.

• Μοντέλα κεντροειδών (Centroid models). Ένα χαρακτηριστικό παράδειγμα είναι και ο αλγόριθμος k-means που αναπαριστά κάθε συστάδα με ένα μόνο μέσο διάνυσμα.

• Μοντέλα κατανομής (Distribution models). Οι συστάδες δημιουργούνται με στατιστικές κατανομές όπως οι πολυμεταβλητές κανονικές κατανομές που χρησιμοποιεί ο γνωστός αλγόριθμος Μεγιστοποίησης προσδοκίας (Expectation Maximization).

• Μοντέλα πυκνότητας (Density models). Για παράδειγμα ο αλγόριθμος DBSCAN ορίζει τις συστάδες σαν πυκνές περιοχές του χώρου των δεδομένων. Οι εφαρμογές της συσταδοποίησης είναι εξαιρετικά πολλές.

Μερικοί τομείς επιστημών είναι:

• Στη Βιολογία και Βιοτεχνολογία. Στην οικολογία φυτών και ζώων. Στη χρονική και χωρική σύγκριση κοινωνιών από οργανισμούς. Στη δημιουργία ομάδων γονιδίων με συσχετισμένα πρότυπα συμπεριφοράς. Στην ανάλυση ακολουθιών για την ομαδοποίηση ομόλογων ακολουθιών σε οικογένειες γονιδίων.

• Στην Ιατρική. Σε ιατρικές απεικονίσεις για τη διαφοροποίηση μεταξύ διαφόρων τύπων ιστών και αίματος σε τρισδιάστατες εικόνες. Στην τμηματοποίηση

IMRT (intensity-modulated radiation therapy) για τη διαίρεση του χάρτη ροής σε διακριτές περιοχές.

• Στην Επιχειρηματικότητα και το Μάρκετινγκ. Για το διαχωρισμό του πληθυσμού των καταναλωτών σε τμήματα και την κατανόηση των διαφόρων ομάδων των καταναλωτών. Επίσης για την ομαδοποίηση των αγορασμένων προϊόντων.

• Στον Παγκόσμιο Ιστό. Στη μελέτη των κοινωνικών δικτύων για την αναγνώριση κοινοτήτων, και στην ομαδοποίηση των διατιθεμένων εγγράφων και ιστοσελίδων για την εύρεση σχετικότερων αποτελεσμάτων από τις μηχανές αναζήτησης.

2.3 Επιλογή του Κατάλληλου Αλγόριθμου

Η επιλογή του κατάλληλου αλγορίθμου για Clustering δεν είναι απλό πράγμα. Η πληθώρα αλγορίθμων Clustering οι οποίοι υπάρχουν στην βιβλιογραφία είναι ένα μεγάλο εμπόδιο στην απόφαση του καλύτερου αλγορίθμου για το εκάστοτε πρόβλημα που αντιμετωπίζεται. Ένα σύνολο από κριτήρια αποδοχής κάποιου αλγορίθμου έχουν προταθεί για την σύγκριση αλγορίθμων Clustering. Αυτά βασίζονται

1) στο τρόπο με τον οποίο σχηματίζονται τα clusters,

2) την δομή που έχουν τα δεδομένα προς επεξεργασία,

3) στην ευαισθησία που έχει ο αλγόριθμος σε αλλαγές που δεν επηρεάζουν τα δεδομένα.

Παράλληλα με αυτά τα κριτήρια θα βοηθούσε πολύ να μπορούσαμε να δώσουμε απαντήσεις και σε άλλα ερωτήματα όπως (Α) ποιο είναι το καλύτερο μέτρο για την σύγκριση της ομοιότητας των στοιχείων, (Β) πως πρέπει να αξιοποιηθεί κάποια γνώση που έχουμε για τα δεδομένα κ.α. Το κυρίως πρόβλημα ενός αλγορίθμου Clustering είναι ότι δεν μπορεί για όλες τις περιπτώσεις δεδομένων να εφαρμοστεί και να αναδείξει επιτυχώς την ποικιλία δομών που εμφανίζονται ειδικά σε πολυδιάστατα σύνολα δεδομένων.

Για τον λόγο αυτό είναι απαραίτητο για κάθε χρήστη ενός αλγορίθμου Clustering να γνωρίζει πολύ καλά την τεχνική που ακολουθεί ο αλγόριθμος της επιλογής του, να έχει γνώση των λεπτομερειών στο πως ομαδοποιούνται τα δεδομένα σε clusters και να είναι καλός γνώστης της πληροφορίας που πρόκειται να επεξεργαστεί. Όσο περισσότερη πληροφορία για τα δεδομένα έχει στα χέρια του ο χρήστης τόσο καλύτερα θα εκτιμηθεί η διαδικασία του Clustering και σωστά

συμπεράσματα θα προκύψουν. Επίσης η γνώση για τα δεδομένα μπορεί να χρησιμοποιηθεί για να βελτιώσει την ποιότητα παραγωγής χαρακτηριστικών, να επιλεγεί το καλύτερο μέτρο ομοιότητας και να αποφασιστεί η όσο το δυνατόν καλύτερη αναπαράσταση των δεδομένων.

2.4 Ορισμοί και Συμβολισμός

Πριν προχωρήσουμε θα δώσουμε έναν αριθμό από ορισμούς και συμβολισμούς που είτε χρησιμοποιήθηκαν είτε θα χρησιμοποιηθούν στην συνέχεια και που στόχο έχουν να ορίσουν πιο φορμαλιστικά κάποιες έννοιες.

- Ένα στοιχείο (pattern) (ή διάνυσμα χαρακτηριστικών) x είναι ένα απλό δεδομένο που επεξεργάζεται από τον αλγόριθμο του Clustering. Αποτελείται από έναν αριθμό d χαρακτηριστικών και συμβολίζεται $x=(x_1,x_2,\ldots,x_d)$.
- Το κάθε μέρος x_i του στοιχείου x καλείται χαρακτηριστικό η γνώρισμα (feature, attribute).
- Ο αριθμός d ορίζει την διάσταση (dimensionality) του κάθε στοιχείου αλλά και την διάσταση του χώρου των δεδομένων.
- Ένα σύνολο από στοιχεία (pattern set) ορίζεται ως εξής $X=\{x_1,x_2,\ldots,x_n\}$. Σε πολλές περιπτώσεις το σύνολο στοιχείων αναπαρίσταται και από έναν πίνακα $n*d$.
- Η κλάση (class) μπορεί να θεωρηθεί σαν μια πηγή στοιχείων τα οποία έχουν κοινά χαρακτηριστικά η όμοια χαρακτηριστικά. Οι αλγόριθμοι Clustering προσπαθούν να δημιουργήσουν σύνολα στοιχείων τα οποία λογικά αναπαριστούν κλάσεις.
- Το μέτρο της απόστασης (distance measure) είναι ένα μέτρο ορισμένο στο χώρο των χαρακτηριστικών των στοιχείων και φανερώνει το πόσο όμοια ή διαφορετικά είναι δύο στοιχεία μεταξύ τους.

2.5 Αναπαράσταση στοιχείων, επιλογή και εξαγωγή χαρακτηριστικών

Όπως σε πολλά ζητήματα που αφορούν το Clustering, έτσι και εδώ δεν υπάρχουν σαφείς οδηγίες, αρχές που να προτείνουν την βέλτιστη αναπαράσταση των στοιχείων και την καλύτερη επιλογή από χαρακτηριστικά. Ο ρόλος του ενδιαφερόμενου στην αναπαράσταση των στοιχείων προς ομαδοποίηση είναι να

συγκεντρώσει όσο το δυνατόν περισσότερα στοιχεία (γνώσεις, συμπεράσματα, υποθέσεις) σχετικά με τα δεδομένα.

Στην συνέχεια με βάση την γνώση για τα δεδομένα την κατάλληλη επιλογή και παραγωγή γνωρισμάτων θα πραγματοποιήσει την βέλτιστη αναπαράσταση για τα στοιχεία που θα επεξεργαστεί ο αλγόριθμος. Υποθέτουμε πάντα ότι η αναπαράσταση των στοιχείων προηγείται της διαδικασίας του Clustering και αυτό είναι φυσικό. Επίσης μια πολύ προσεκτική ματιά στα σημαντικότερα χαρακτηριστικά και πιθανόν κάποιοι μετασχηματισμοί πάνω σε αυτά θα μπορούσαν να οδηγήσουν σε καλύτερα αποτελέσματα. Μια καλή αναπαράσταση των στοιχείων συνήθως οδηγεί σε μια απλή και κατανοητή ομαδοποίηση. Αντίθετα μια όχι και τόσο προσεκτική αναπαράσταση μπορεί να οδηγήσει σε μια πολύπλοκη ομαδοποίηση με δομές δύσκολα αντιληπτές και εκτιμησημες.

Κάθε στοιχείο με την βοήθεια των χαρακτηριστικών του προσδιορίζει ένα φυσικό αντικείμενο ή μια αφηρημένη έννοια. Κάθε χαρακτηριστικό ενός στοιχείου, το οποίο είναι ένα πολυδιάστατο διάνυσμα, αντιπροσωπεύει και ποσοτικοποιεί μια από τις d διαστάσεις μέσα στις οποίες υπάρχει το στοιχείο. Τα χαρακτηριστικά μπορεί να εκφράζουν ποσότητα η ποιότητα. Για παράδειγμα αν το ύψος και το χρώμα είναι δύο χαρακτηριστικά που χρησιμοποιούνται για να περιγράψουν ένα στοιχείο, τότε το διάνυσμα (20, μαύρο) είναι η αναπαράσταση ενός στοιχείου χρώματος μαύρου και ύψους 20 μονάδων. Τα χαρακτηριστικά μπορούν να χωριστούν σε διάφορους τύπους:

- Ποσοτικά χαρακτηριστικά:
  - ο Συνεχών τιμών (βάρος, ύψος)
  - ο Διακριτών τιμών (αριθμός ατόμων, τηλεοράσεων)
  - ο Τιμές διαστημάτων (διάρκεια κάποιου γεγονότος)

- Ποιοτικά χαρακτηριστικά:
  - ο Ονομαστικά ή μη ταξινομήσιμα (χρώμα)
  - ο Ταξινομήσιμα (θερμοκρασία – ζεστό, κρύο - , θόρυβος – σιγά, δυνατά)

Τις περισσότερες φορές είναι χρήσιμο να απομονώνουμε εκείνα τα χαρακτηριστικά που διαφαίνονται να είναι πιο σημαντικά, πιο περιγραφικά, που προσφέρουν μεγαλύτερη διακριτική ικανότητα στον αλγόριθμο. Η διαδικασία

επιλογής χαρακτηριστικών (feature selection) στόχο έχει την εύρεση ενός υποσυνόλου χαρακτηριστικών τα οποία και τελικά θα χρησιμοποιηθούν. Αντίθετα η τεχνική της παραγωγής χαρακτηριστικών στοχεύει στην δημιουργία νέων χαρακτηριστικών από τα ήδη υπάρχοντα. Και οι δυο τεχνικές αποσκοπούν στην βελτίωση της κατηγοριοποίησης και την καλύτερη υπολογιστική απόδοση.

2.6 Μέτρο ομοιότητας

Σε κάθε cluster τα στοιχεία που περιέχονται σε αυτό παρουσιάζουν ομοιότητα μεταξύ τους και αυτό είναι βασικό για να ορισθεί ένα ξεχωριστό cluster. Έτσι για όλες τις τεχνικές Clustering είναι σημαντικό να ορίζεται ένα μέτρο ομοιότητας μεταξύ δύο στοιχείων από το χώρο δεδομένων. Δεδομένης της μεγάλης ποικιλίας στα χαρακτηριστικά των στοιχείων η επιλογή του μέτρου ομοιότητας θα πρέπει να είναι πολύ προσεγμένη. Σε πολλές περιπτώσεις με το μέτρο ομοιότητας αυτό που συνήθως μετράται δεν είναι η ομοιότητα αλλά η διαφορετικότητα δυο τυχαίων στοιχείων όπως έχουμε ήδη αναφέρει. Στην συνέχεια θα αναφερθούμε σε μέτρα ομοιότητας τα οποία είναι ευρέως διαδεδομένα, και χρησιμοποιούνται για την σύγκριση στοιχείων των οποίων τα χαρακτηριστικά περιγράφονται από συνεχείς τιμές. Το μέτρο ομοιότητας καλείται και απόσταση και ικανοποιεί την τριγωνική ανισότητα για δύο στοιχεία x,y:

D(x,y) = 0 αν και μόνο αν x = y

D(x,y) = D(y,x)

D(x,z) ≤ D(x,y) + D(y,z)

Το πιο γνωστό μέτρο ομοιότητας που χρησιμοποιείται είναι η Ευκλείδεια απόσταση η οποία ορίζεται ως εξής:

$$D(x,y) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Άλλοι τύποι που δίνουν την απόσταση μεταξύ δύο στοιχείων μπορεί να είναι η απόσταση Manhattan:

$$\mathbf{D(x,y)} = \sum_{i=1}^{k} | x_i - y_i |$$

ή το μέγιστο της διαφοράς σε κάθε διάσταση:

$$\mathbf{D(x,y)} = \max_{i=1}^{k} | x_i - y_i |$$

Η ευκλείδεια απόσταση χρησιμοποιείται ευρέως σε περιπτώσεις λίγων διαστάσεων και έχει καλά αποτελέσματα όταν δεδομένα κατηγοριοποιούνται σε συμπαγή και αρκετά απομονωμένα clusters. Ένα πρόβλημα που παρουσιάζει είναι ότι στις πολλές διαστάσεις το χαρακτηριστικό το οποίο παρουσιάζει την μεγαλύτερη διαφοροποίηση από τα άλλα κυριαρχεί και αποπροσανατολίζει το τελικό αποτέλεσμα. Εδώ πρόκειται για αυτό που συνήθως αναφέρεται ως κατάρα των πολλών διαστάσεων (curse of dimensionality).

Μερικοί αλγόριθμοι αντί να υπολογίζουν κάθε φορά την απόσταση μεταξύ δύο στοιχείων, χρησιμοποιούν ένα πίνακα στον οποίο τοποθετούν τις ομοιότητες των στοιχείων. Αυτό που γίνεται είναι ένας προ-υπολογισμός των n(n-1)/2 τιμών ομοιότητας για ένα σύνολο n στοιχείων. Όσον αφορά τώρα τον υπολογισμό της απόστασης για στοιχεία των οποίων τα χαρακτηριστικά δεν είναι συνεχείς τιμές , αυτός είναι αρκετά προβληματικός.

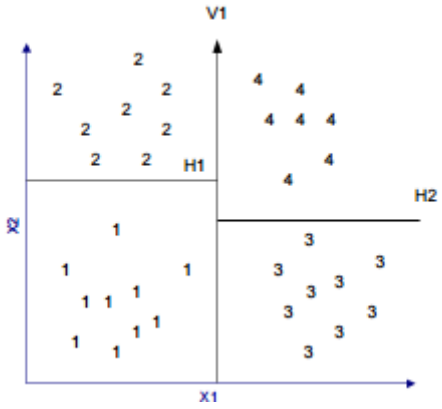
Στις περισσότερες των περιπτώσεων τα χαρακτηριστικά δεν είναι συγκρίσιμα και το αποτέλεσμα της σύγκρισης έχει δύο δυνατές τιμές, όμοιο ή ανόμοιο. Παρόλα αυτά οι ειδικοί που ασχολούνται με στοιχεία των οποίων τα χαρακτηριστικά είναι και των δύο τύπων έχουν βρει μεθόδους και μέτρα για τον ορισμό της απόστασης των στοιχείων

2.7 Τεχνικές Clustering

Οι τεχνικές Clustering μπορούν να διαχωριστούν με πολλούς τρόπους, όπως Ιεραρχικό Clustering σε αντίθεση με το Διαμεριστικό, και καθένα από αυτά να χωριστεί σε άλλες υποκατηγορίες. Θα αναφερθούμε σε διαφορετικές προσεγγίσεις Clustering παρακάτω αφού προηγουμένως δούμε κάποιους όρους και διαφοροποιήσεις που παρουσιάζουν οι διάφοροι αλγόριθμοι. Οι αλγόριθμοι για Clustering μπορεί να είναι:

- Συγκεντρωτικοί και Διαχωριστικοί (Agglomerative and Divisive). Η διαφοροποίηση των ειδών αυτών σχετίζεται με την λειτουργία και τις δομές του αλγορίθμου. Στην πρώτη περίπτωση ο αλγόριθμος ξεκινά θεωρώντας κάθε στοιχείο

σαν ένα ξεχωριστό cluster, και προχωρά συγχωνεύοντας στοιχεία και clusters μέχρις ότου να ικανοποιηθεί μια συνθήκη. Στην περίπτωση ενός διαχωριστικού αλγορίθμου, όλα τα στοιχεία θεωρούνται ότι ανήκουν σε ένα cluster και ακολουθείται μια συνεχής διάσπαση του cluster αυτού σε υπό- cluster μέχρις ότου να ικανοποιηθεί η συνθήκη τερματισμού.



Πηγή: Nigam et al., 2000

- Μονοθετικοί και Πολυθετικοί (Monothetic and Polythetic). Η διαφορά αυτών χαρακτηρίζει την σειριακή ή ταυτόχρονη χρησιμοποίηση των χαρακτηριστικών των στοιχείων κατά την διαδικασία του Clustering. Οι περισσότεροι αλγόριθμοι είναι πολυθετικοί, κάτι που σημαίνει ότι όλα τα χαρακτηριστικά των στοιχείων συμμετέχουν κάθε φορά στον καθορισμό της απόστασης του στοιχείου από κάποιο άλλο. Ένας μονοθετικός αλγόριθμος λαμβάνει υπόψη του μόνο ένα χαρακτηριστικό τη φορά και πραγματοποιεί ομαδοποιήσεις με βάση αυτό το χαρακτηριστικό. Σε επόμενη επανάληψη χρησιμοποιεί άλλο χαρακτηριστικό και διαχωρίζει τις ήδη υπάρχουσες ομάδες. Ένα παράδειγμα φαίνεται στο σχήμα 3.3. Εδώ τα στοιχεία του χώρου μας έχουν χωριστεί σε δύο clusters αρχικά με βάση το χαρακτηριστικό X1. Ο διαχωρισμός δηλώνεται με την κάθετη γραμμή V. Στην συνέχεια κάθε cluster χωρίζεται με βάση το χαρακτηριστικό X2 και τα νέα clusters διαχωρίζονται από τις οριζόντιες γραμμές H1 και H2. Το πρόβλημα αυτών των αλγορίθμων είναι ότι τα στοιχεία χωρίζονται τελικά σε 2d clusters όπου d είναι ο αριθμός των χαρακτηριστικών των στοιχείων. Αυτό συνήθως οδηγεί σε πολλά clusters εκ των οποίων τα περισσότερα είναι μικρά και ασήμαντα.

- Σκληροί και fuzzy (hard and fuzzy). Ένας σκληρός αλγόριθμος τοποθετεί κάθε στοιχείο σε ένα και μόνο cluster, σε αντίθεση με τους fuzzy αλγορίθμους οι

οποίοι δίνουν σε κάθε στοιχείο για κάθε cluster έναν βαθμό που εκφράζει κατά πόσο το στοιχείο αυτό ανήκει στο cluster αυτό.

- Ντετερμινιστικοί και Στοχαστικοί (Deterministic and Stochastic). Αυτοί οι αλγόριθμοι είναι κυρίως διαιρετικοί και σχετίζονται με την βελτιστοποίηση της ομαδοποίησης.

- Αυξηντικοί και μη αυξηντικοί (Incremental and non-incremental). Η διαφορά αυτών των αλγορίθμων εμφανίζεται όταν το σύνολο των δεδομένων προς ομαδοποίηση είναι πολύ μεγάλο και περιορισμοί που υπάρχουν στον χρόνο εκτέλεσης και τον διαθέσιμο χώρο μνήμης επηρεάζουν την αρχιτεκτονική του αλγορίθμου. Στα πρώτα βήματα της θεωρίας περί clustering τα δεδομένα δεν ήταν ιδιαίτερα πολλά και προβλήματα με το μέγεθος της πληροφορίας δεν υπήρχαν. Με την αύξηση όμως της πληροφορίας υπήρξε η ανάγκη για εύρεση αλγορίθμων οι οποίοι ελαχιστοποιούν τον αριθμό σαρώσεων των δεδομένων, μειώνουν τον αριθμό των στοιχείων που εξετάζονται ή μειώνουν το μέγεθος των δομών που χρησιμοποιούνται κατά την εκτέλεση του αλγορίθμου.

2.8 Expectation Maximization (EM)

Στη στατιστική ένας αλγόριθμος μεγιστοποίησης προσδοκίας είναι μια επαναληπτική διαδικασία για την εύρεση της μέγιστης πιθανότητας ή μέγιστες εκ των υστέρων (maximum a posteriori – MAP) εκτιμήσεις των παραμέτρων σε στατιστικά μοντέλα, όπου το μοντέλο βασίζεται σε μη φανερές μεταβλητές.

Ο αλγόριθμος ΕΜ συνέτεινε στην εξαγωγή πολύ σημαντικών θεωρητικών και πρακτικών αποτελεσμάτων. Από τη μεριά των θεωρητικών προσπαθειών, μεγάλο μέρος δαπανήθηκε στον προσδιορισμό του πλήθους των μη ταξινομημένων παραδειγμάτων που αξίζουν όσο ένα ταξινομημένο. Ένα από τα πλέον γνωστά θεωρητικά αποτελέσματα συνοψίζεται ως ακολούθως: Βάσει ορισμένων υποθέσεων (Joachims, 1999):

• εφόσον έχουμε στη διάθεσή μας άπειρο αριθμό από μη ταξινομημένα παραδείγματα, τα ταξινομημένα παραδείγματα που διατίθενται ελαττώνουν το σφάλμα ταξινόμησης εκθετικά γρήγορα

• αν διατίθεται πεπερασμένος αριθμός από μη ταξινομημένα παραδείγματα, τα ταξινομημένα είναι εκθετικά πιο πλούσια σε πληροφοριακό περιεχόμενο απ' ότι τα πρώτα.

Στον τομέα της εφαρμοσμένης έρευνας, ο αλγόριθμος ΕΜ με χρήση ταξινομημένων και μη παραδειγμάτων βρήκε εφαρμογή σε μια ποικιλία πεδίων, όπως αυτά της αναγνώρισης προσώπων και της ταξινόμησης κειμένου και συνδυάστηκε με αρκετούς αλγορίθμους μάθησης, όπως ο Naive Bayes και τα δένδρα εξάρτησης

Στα πλαίσια της εργασίας των Nigam et al. (2000), παρατίθεται ο βασικός αλγόριθμος της προσέγγισής τους, που περιλαμβάνει τη χρήση του ΕΜ για την εκμάθηση των παραμέτρων ενός συνόλου n μοντέλων παραγωγής από ταξινομημένα και μη κείμενα (ήτοι ενός ταξινομητή n κλάσεων). Ο αλγόριθμος, που σκιαγραφείται στο παρακάτω σχήμα, περιλαμβάνει την επαναληπτική εκτέλεση δύο βημάτων. Στο πρώτο εξ αυτών, τα άγνωστα παραδείγματα ταξινομούνται από τα τρέχοντα μοντέλα παραγωγής και για κάθε παράδειγμα εξάγεται ένας βαθμός εμπιστοσύνης για κάθε κλάση (που αντιστοιχεί στην πιθανότητα το παράδειγμα να ανήκει στη δεδομένη κλάση ή όχι). Στο δεύτερο βήμα, οι παράμετροι των μοντέλων παραγωγής επανεκτιμούνται (δηλ. ο ταξινομητής επανεκπαιδεύεται), χρησιμοποιώντας τους βαθμούς εμπιστοσύνης κάθε παραδείγματος από το προηγούμενο βήμα.

Βασικός αλγόριθμος ΕΜ (Nigam et al., 2000)

Είσοδος: το αρχικό σύνολο ταξινομημένων παραδειγμάτων D_l, τα μη ταξινομημένα παραδείγματα D_u.

Έξοδος: ο ταξινομητής Τ' που εκτιμά την κλάση αγνώστων παραδειγμάτων. Εκπαίδευσε τον αρχικό ταξινομητή Τ με τα προταξινομημένα παραδείγματα του συνόλου D_l. Όσο οι παράμετροι του ταξινομητή βελτιώνονται:

• Βήμα-Ε: Εκτίμησε για κάθε μη ταξινομημένο παράδειγμα την πιθανότητα να ανήκει σε κάθε μια από τις κλάσεις του προβλήματος, χρησιμοποιώντας το τρέχοντα ταξινομητή Τ.

• Βήμα-Μ: Επανεκπαίδευσε τον ταξινομητή Τ, λαμβάνοντας υπ' όψη τις πιθανότητες του βήματος-Ε για κάθε μη ταξινομημένο παράδειγμα

Θα πρέπει να επισημανθεί ότι οι προσεγγίσεις μάθησης με μερική επίβλεψη που βασίζονται στον ΕΜ δίνουν ικανοποιητικά αποτελέσματα μόνο όταν οι υποθέσεις που έχουμε κάνει για τις κατανομές που ακολουθούν τα δεδομένα του μοντέλου παραγωγής είναι συνεπείς με αυτές στον πραγματικό κόσμο. Μάλιστα, έχει αποδειχθεί πως αν η παραπάνω προϋπόθεση δεν ισχύει, τα μη ταξινομημένα δεδομένα ενδέχεται να βλάψουν την απόδοση του τελικού ταξινομητή. Σε αυτήν την περίπτωση, όταν δηλαδή το μοντέλο παραγωγής δεν είναι κατάλληλο, έχει προταθεί ο αλγόριθμος M-EM [NMTM00] που χρησιμοποιεί πολλαπλές υποκατηγορίες για κάθε κλάση, το πλήθος των οποίων καθορίζεται από διασταυρωμένη επικύρωση πάνω στο σώμα εκπαίδευσης. Ένας άλλος αλγόριθμος που αποδίδει πολύ καλύτερα από τον M-EM είναι ο Partitioned-EM [CLWL04] που χρησιμοποίει ιεραρχική ομαδοποίηση.

2.9 Ο αλγόριθμος k-means

Ο αλγόριθμος αυτός είναι μια μέθοδος διανυσματικού κβαντισμού και είναι πολύ δημοφιλής στο κόσμο της εξόρυξης γνώσης. Χρησιμοποιείται στο διαχωρισμό των δεδομένων σε k συστάδες όπου κάθε σημείο δεδομένων ανήκει στη συστάδα με τον κοντινότερο μέσο όρο ο οποίος αποτελεί και αντιπροσωπευτικό σημείο της συστάδας. Ο μέσος αυτός όρος ονομάζεται κεντροειδής (centroid) της συστάδας. Ο χωρισμός αυτός των δεδομένων γίνεται στα λεγόμενα κελιά Voronoi. Το πρόβλημα είναι υπολογιστικά δύσκολο αλλά μπορεί να λυθεί με ευριστικούς αλγορίθμους που συγκλίνουν γρήγορα σε τοπικό ελάχιστο. Αυτοί μοιάζουν με τον αλγόριθμο ΕΜ που χρησιμοποιείται σε μίγματα γκαουσιανών κατανομών.

Επίσης και ο ΕΜ και ο k-means χρησιμοποιούν κεντροειδείς συστάδων για τη μοντελοποίηση αλλά στον k-means οι συστάδες τείνουν να έχουν παρόμοιο χωρικό εύρος ενώ στον ΕΜ μπορούν να έχουν διαφορετικά σχήματα. Η εύρεση των συστάδων στον k-means γίνεται με την ελαχιστοποίηση του κριτηρίου που είναι γνωστό ως αδράνεια (inertia) ή με το άθροισμα των τετραγώνων των αποστάσεων των εντός της συστάδας δεδομένων. Απαιτεί να του ορίσουμε τον αριθμό των συστάδων σαν παράμετρο ενώ κλιμακώνεται καλά σε μεγάλη ποσότητα δεδομένων. Ο τρόπος υπολογισμού της αδράνειας χρησιμοποιεί την Ευκλείδεια απόσταση αλλά σε χώρους πολλών διαστάσεων η απόσταση αυτή τείνει να παραφουσκώσει, ένα χαρακτηριστικό που ονομάζεται “κατάρα των διαστάσεων”.

Σε βασικούς όρους ο αλγόριθμος έχει τρία βήματα. Στο πρώτο επιλέγονται οι αρχικοί κεντροειδείς, με την πιο απλή μέθοδο επιλογής να είναι k δείγματα από τα δεδομένα. Μετά την αρχικοποίηση ο αλγόριθμος ανακυκλώνεται στα επόμενα δυο βήματα. Στο πρώτο αντιστοιχίζει κάθε δεδομένο στον κοντινότερό του κεντροειδή. Στο δεύτερο βήμα δημιουργεί νέους κεντροειδείς από το μέσο όρο όλων των δεδομένων που είχαν αντιστοιχηθεί στον προηγούμενο κεντροειδή. Τέλος υπολογίζεται η απόσταση του παλιού με το νέο κεντροειδή και τα βήματα αυτά επαναλαμβάνονται μέχρι η απόσταση να γίνει μικρότερη από ένα κατώφλι, δηλαδή μέχρις ότου οι κεντροειδείς δεν μετακινούνται σημαντικά.

Μετά από ένα αριθμό επαναλήψεων ο k-means πάντα συγκλίνει αλλά μπορεί αυτό να γίνει σε τοπικό ελάχιστο. Στο αποτέλεσμα αυτό παίζει σημαντικό ρόλο η αρχικοποίηση των κεντροειδών. Έτσι πολλές φορές ο αλγόριθμος επαναλαμβάνεται μερικές φορές με διαφορετικούς αρχικούς κεντροειδείς. Μια μέθοδος που ξεπερνά αυτό το πρόβλημα και υποστηρίζεται στη βιβλιοθήκη scikit-learn είναι το σχήμα αρχικοποίησης k-means++. Σε αυτό η αρχικοποίηση των κεντροειδών γίνεται σε μεγάλη απόσταση ο ένας από τον άλλο οδηγώντας σε καλύτερα αποτελέσματα από την τυχαία αρχικοποίηση.

Στην συνέχεια περιγράφονται με την μορφή ψευδοκώδικα τα βασικά βήματα του αλγορίθμου k-means (Kanungo et al., 2002). Ο αλγόριθμος ξεκινά καθορίζοντας με τυχαίο τρόπο τα κέντρα που θα αντιπροσωπεύουν τις c συστάδες. Στην συνέχεια προσδιορίζεται η απόσταση κάθε στοιχείου του συνόλου δεδομένων από το κέντρο κάθε συστάδας και κάθε στοιχείο τοποθετείται στην συστάδα από την οποία απέχει λιγότερο. Τα κέντρα των νέων συστάδων υπολογίζονται σαν ο μέσος όρος των στοιχείων που ανήκουν μέχρι στιγμής σε κάθε συστάδα. Η διαδικασία επαναλαμβάνεται μέχρις ότου οι συστάδες να σταματήσουν να μεταβάλλονται. Αυτό σημαίνει ότι η απόκλιση μεταξύ των κέντρων των συστάδων που προέκυψαν τελευταία από αυτά της προηγούμενης επανάληψης είναι κοντά στο μηδέν (τα κέντρα ταυτίζονται). Τα βήματα του αλγορίθμου σε μορφή ψευδοκώδικα είναι τα εξής:

1. Εύρεση των αρχικών κέντρων, v_i i = 1,2,..., c, για τις c συστάδες. Για κάθε επανάληψη r = 1,..., r_{max}

2. Υπολογισμός της απόστασης κάθε στοιχείου του συνόλου δεδομένων από το κέντρο κάθε συστάδας

$$d_{ki} = (x_k - v_i)2,\ k=1,2,\ldots,n\ i=1,2,....,c$$

3. Κάθε στοιχείο χ_κ αντιστοιχίζεται στην συστάδα για την οποία ισχύει

$$Min_{k,i}(d_{ik}),\ \forall\ i,k$$

4. Υπολογισμός των νέων κέντρων των συστάδων

$$m_i^{(r+1)} = \frac{\sum_{k=1}^{n_i} x_k}{n_i}$$

όπου n_i , ο αριθμός των στοιχείων που ανήκουν στην i συστάδα μέχρι στιγμής.

5. If || mi(r) – mi(r+1) || < ε then stop
   else
   r=r+1, goto2.

Ο k-means όπως προαναφέρθηκε αποτελεί μία ευρέως αποδεκτή τεχνική συσταδοποίησης, η οποία έχει χρησιμοποιηθεί αποτελεσματικά για συσταδοποίηση σε διάφορα πεδία ορισμού. Ωστόσο, ο k-means δεν είναι η μοναδική τεχνική, υπάρχουν διάφορες εκδόσεις και πλήθος παραλλαγών αυτής. Οι παραλλαγές αυτές διαφέρουν κυρίως στον τρόπο επιλογής των αρχικών k μέσων (κέντρων) των συστάδων, στον υπολογισμό της ομοιότητας και στη (στρατηγική που χρησιμοποιούν για τον υπολογισμό των μέσων των συστάδων. Επιπρόσθετα, διάφορα στατιστικά πακέτα όπως το SAS, SPSS και BMPD που χρησιμοποιούν τον k-means υιοθετούν την δική τους έκδοση το καθένα για τον αλγόριθμο (Kanungo et al., 2002).

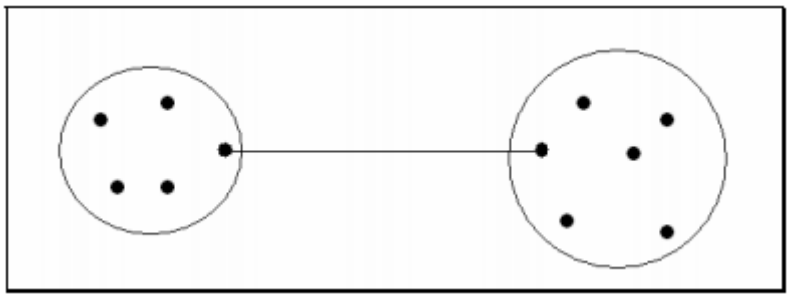
2.10 Ιεραρχικοί αλγόριθμοι ομαδοποίησης

Οι ιεραρχικοί αλγόριθμοι δεν διαιρούν τα δεδομένα σε συγκεκριμένο αριθμό ομάδων σε ένα μόνο βήμα. Αντίθετα, η διαίρεση σε ομάδες περιλαμβάνει πολλαπλά βήματα στα οποία δημιουργείται μια σειρά από διαμερίσεις του αρχικού συνόλου. Η σειρά αυτή αρχίζει από μια μόνο συστάδα με όλα μαζί τα δεδομένα και καταλήγει σε n συστάδες με κάθε μια να αποτελείται από ένα μόνο αντικείμενο. Στην αντίστροφη

περίπτωση ξεκινάμε με n συστάδες του ενός στοιχείου και καταλήγουμε σε μία συστάδα με όλα τα στοιχεία. Έτσι, διακρίνουμε δύο κατηγορίες μεθόδων, τις μεθόδους συσσώρευσης και τις μεθόδους διαίρεσης.

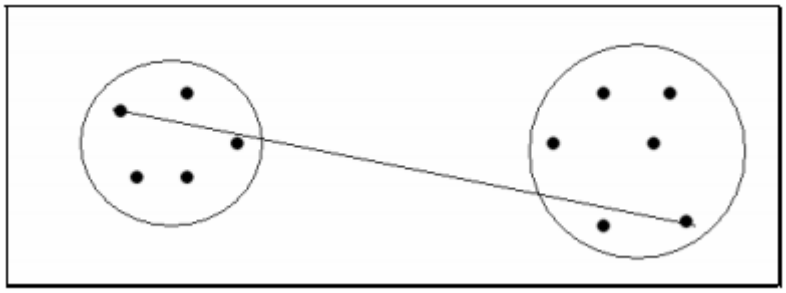
Στις μεθόδους συσσώρευσης δημιουργείται μία σειρά από διαδοχικές συγχωνεύσεις των n αντικειμένων σε ομάδες ενώνοντας κάθε φορά 2 ομάδες που βρίσκονται πιο κοντά με βάση την ομοιότητα η την απόσταση μεταξύ τους. Αρχικά, όταν κάθε ένα από τα αντικείμενα αποτελεί μια συστάδα, η έννοια της απόστασης των συστάδων συμπίπτει με την απόσταση των αντικειμένων, αλλά στη συνέχεια όταν οι συστάδες αποτελούνται πλέον από περισσότερα αντικείμενα, είναι βασικό να καθοριστεί ο τρόπος μέτρησης της απόστασης μεταξύ τους. Οι βασικές μέθοδοι υπολογισμού την απόστασης μεταξύ δυο ομάδων που χρησιμοποιούν οι συσωρευτικοί ιεραρχικοί αλγόριθμοι είναι (Oikonomakou et al., 2005):

- Μέθοδος απλού συνδέσμου (single linkage): Ως απόσταση ανάμεσα στις συστάδες C_i, C_j ορίζεται η μικρότερη από τις αποστάσεις ανάμεσα σε κάθε στοιχείο της ομάδας C_i και κάθε στοιχείο της ομάδας C_j. Στο σχήμα που ακολουθεί φαίνεται ο τρόπος υπολογισμού της απόστασης .



Πηγή: Nigam et al., 2000

- Μέθοδος του πλήρους συνδέσμου (complete linkage): Ως απόσταση ανάμεσα στις συστάδες C_i, C_j, ορίζεται η μεγαλύτερη από τις αποστάσεις ανάμεσα σε κάθε στοιχείο της ομάδας C_i και κάθε στοιχείο της ομάδας C_j. Στο σχήμα που ακολουθεί φαίνεται ο τρόπος υπολογισμού της απόστασης .



Πηγή: Nigam et al., 2000

- Μέθοδος της μέσης απόστασης (group average or average linkage method): Στην περίπτωση αυτή ορίζεται ως απόσταση δύο συστάδων, η μέση τιμή όλων των αποστάσεων ανάμεσα σε κάθε σημείο της ομάδας C_i και κάθε σημείο της ομάδας C_j. Θεωρείται «ενδιάμεση» μέθοδος ανάμεσα στις μεθόδους απλού συνδέσμου και πλήρους συνδέσμου. Το πλεονέκτημα της μεθόδου είναι ότι λαμβάνει υπόψη τη δομή των συστάδων. □

- Μέθοδος του κεντροειδούς (centroid linkage): Στην περίπτωση αυτή για τον υπολογισμό της απόστασης των συστάδων χρησιμοποιούνται δύο "αντιπροσωπευτικά" σημεία, τα κέντρα των ομάδων, τα οποία υπολογίζονται από όλα τα αντικείμενα της κάθε ομάδας.

Και οι δυο κατηγορίες ιεραρχικής ταξινόμησης βασίζονται στη δομή του πίνακα εγγύτητας και χρησιμοποιούν κάποιο κριτήριο βελτιστοποίησης της διαδικασίας ένωσης ή υποδιαίρεσης των ομάδων. Ένα βασικό χαρακτηριστικό, που ταυτόχρονα θα μπορούσε να θεωρηθεί και ως μειονέκτημα των μεθόδων αυτών, είναι ότι οι συγχωνεύσεις ή οι υποδιαιρέσεις που γίνονται κατά τη διάρκεια της διαδικασίας, είναι ανεπανόρθωτες. Αυτό σημαίνει ότι από τη στιγμή που κάποια αντικείμενα ενωθούν, θα μείνουν μαζί μέχρι το τέλος της διαδικασίας και αντίστοιχα αν χωριστούν θα μείνουν χωρισμένα.

Ο γενικός αλγόριθμος της ιεραρχικής συσταδοποίησης παρουσιάζεται στη συνέχεια. Γενικά υπάρχουν πολλές τεχνικές ιεραρχικής συσταδοποίησης που είναι παραλλαγές μιας συγκεκριμένης προσέγγισης: ξεκινώντας με διαφορετικά σημεία

σαν συστάδες, επιτυχώς ενώνονται οι δυο πιο κοντινές συστάδες μέχρι να μείνει μόνο μια συστάδα. Αυτή η προσέγγιση αναλύεται παρακάτω:

1. Υπολογισμός του πίνακα γειτνίασης, εάν είναι απαραίτητο

2. Επανάληψη

3. Ένωση των δυο πιο κοντινών συστάδων

4. Ενημέρωση της εγγύτητας του πίνακα για ανανέωση της εγγύτητας μεταξύ της νεάς συστάδας και των πραγματικών συστάδων

5. Μέχρι μόνο μια συστάδα να μείνει.

Μερικοί αντιπροσωπευτικοί ιεραρχικοί αλγόριθμοι είναι: ο αλγόριθμος BIRCH, (Zhang et al., 1996), ο οποίος είναι κατάλληλος για πολύ μεγάλα σύνολα δεδομένων, μπορεί να διαχειρίζεται τα ακραία σημεία αποτελεσματικά και μπορεί να λειτουργήσει με περιορισμένη κύρια μνήμη. Ο αλγόριθμος CURE (Guha et al., 1998) μπορεί να εφαρμοστεί αποδοτικά και για ομαδοποίηση μεγάλων βάσεων δεδομένων συνδυάζοντας τεχνικές τυχαίας δειγματοληψίας (sampling) και τμηματοποίησης (partitioning) και μπορεί να χειρίζεται περιορισμένη μνήμη. Ο αλγόριθμος ROCK (Guha et al., 1999), ο οποίος μπορεί να διαχειριστεί δυαδικά και κατηγορικά γνωρίσματα.

2.11 Ομαδοποίηση βασισμένη-στην-πυκνότητα (Density-based clustering)

Ο DBSCAN (Sander et al., 1998) είναι ένας απλός και αποδοτικός αλγόριθμος συσταδοποίησης βασιζόμενος στην πυκνότητα και πραγματοποιεί μια σειρά από σημαντικές εργασίες. Γενικά δεν υπάρχουν πολλές μέθοδοι που να βασίζονται στην πυκνότητα και οι περισσότερες εστιάζουν στην ομοιότητα των στοιχείων. Η πιο κλασική περίπτωση μέτρησης πυκνότητας είναι αυτή που βασίζεται στο κέντρο

(center based). Σύμφωνα με αυτή η πυκνότητα εκτιμάται για ένα συγκεκριμένο σημείο του συνόλου δεδομένων με τη μέτρηση του αριθμού των σημείων που βρίσκονται μέσα σε κάποια συγκεκριμένη ακτίνα, Eps, του σημείου αυτού. Αυτός ο αριθμός περιλαμβάνει και το σημείο αυτό. Σε αυτή την προσέγγιση βασίζεται και ο DBSCAN. Η παραπάνω προσέγγιση της πυκνότητας επιτρέπει την κατηγοριοποίηση ενός σημείου (1) που βρίσκεται εσωτερικά μιας πυκνής περιοχής (core point), (2) στην άκρη μιας πυκνής περιοχής (border point), ή (3) σε μια αραιή περιοχή (noise or background point). Παρακάτω δίνεται μια καλύτερη ανάλυση των παραπάνω εννοιών.

- Core points: Αυτά τα σημεία βρίσκονται στο εσωτερικό μιας συστάδας βασισμένης στην πυκνότητα. Ένα σημείο είναι ο πυρήνας της συστάδας εάν ο αριθμός των σημείων μέσα σε μια καθορισμένη γειτονιά γύρω από το σημείο αυτό που βρίσκεται από την συνάρτηση απόστασης και μιας παραμέτρους απόστασης από το χρήστη, Eps, υπερβαίνει κάποιο όριο που επίσης καθορίζεται από το χρήστη.

- Border points: Ένα τέτοιο σημείο δεν είναι πυρήνας όπως πριν αλλά βρίσκεται στη γειτονιά ενός σημείου που είναι πυρήνας. Επίσης μπορεί να βρίσκεται στη γειτονιά πολλών σημείων που είναι πυρήνες.

- Noise points: Τα σημεία εδώ είναι οποιαδήποτε σημεία δεν είναι πυρήνες ή βρίσκονται σε γειτονιές σημείων που είναι πυρήνες.

Αφού αναφέρθηκαν κάποιοι χρήσιμοι ορισμοί ο αλγόριθμος DBSCAN μπορεί να περιγραφεί ως ακολούθως. Οποιαδήπτε δυο core points που είναι αρκετά κοντά σε μια απόσταση Eps το ένα από το άλλο τοποθετούνται στην ίδια συστάδα. Παρόμοια, κάθε border point που είναι αρκετά κοντά σε ένα core point τοποθετείται στην ιδια συστάδα με το core point. Τα noise points αποβάλλονται. Κάποιες λεπτομέριες του αλγορίθμου φαίνονται παρακάτω:

1. Βάζει ετικέτες σε όλα τα σημεία σαν να είναι core, border ή noise points.

2. Αποβάλλει τα noise points.

3. Τοποθετεί μια ακμή μεταξύ όλων των core points που έχουν απόσταση μεταξύ τους Eps.

4. Κάνει κάθε ομάδα των συνδεδεμένων core points μια χωριστή συστάδα.

5. Αντιστοιχίζει κάθε border point σε μια από τις συστάδες των αντίστοιχων core points.

ΚΕΦΑΛΑΙΟ 3

Εμπειρικό Μέρος

3.1 Περιγραφική των Δεδομένων

Τα δεδομένα που συλλέχτηκαν αφορούσαν την περίοδο 1/7/2013-31/12/2013 (*Site Usage*) και αφορούσαν την επισκεψιμότητα της ιστοσελίδας μιας ασφαλιστικής εταιρείας. Από την ανάλυση των δεδομένων που συλλέχτηκαν διαπιστώθηκε πως ο μέσος αριθμός επισκεπτών για το συγκεκριμένο εξάμηνο ήταν 7.486,35 επισκέπτες (Τ.Α.= 3.231,159 επισκέπτες). Το εύρος των επισκεπτών είναι 15.262, γεγονός που σημαίνει ότι υπάρχει μεγάλη μεταβλητότητα στις επισκέψεις της ιστοσελίδας. Οι μοναδικοί επισκέπτες που μετρήθηκαν με βάση την μοναδική IP, η οποία είναι η ηλεκτρονική ταυτότητα του κάθε επισκέπτη είναι κατά μέσο όρο 6.500,06 επισκέπτες (Τ.Α.= 2.926,37 επισκέπτες). Ο συνολικός αριθμός σελίδων που επισκέφτηκαν οι επισκέπτες είναι κατά μέσο όρο 26.755,12 (Τ.Α.=11.446,68 επισκέπτες). Ο μέσος όρος σελίδων που επισκέπτεται ένας επισκέπτης είναι 3.587.

***Πίνακας 1**: Περιγραφικά μέτρα των χαρακτηριστικών μεγεθών που αφορούν την επισκεψιμότητα της ιστοσελίδας*

Μεταβλητή	***Στατιστικό μέτρο***	
visitors	*Μέση τιμή*	7486.350
	Διάμεσος	7342.000
	Τυπική Απόκλιση	3231.159
	Ελάχιστη Τιμή	2216.000
	Μέγιστη Τιμή	17478.000
unique visitors	*Μέση τιμή*	6500.060
	Διάμεσος	6464.000
	Τυπική Απόκλιση	2926.372
	Ελάχιστη Τιμή	1849.000
	Μέγιστη Τιμή	15822.000
pageviews	*Μέση τιμή*	26755.120
	Διάμεσος	28320.000
	Τυπική Απόκλιση	11446.684
	Ελάχιστη Τιμή	6182.000
	Μέγιστη Τιμή	49875.000
pages per visit	*Μέση τιμή*	3.587
	Διάμεσος	3.450

	Τυπική Απόκλιση	0.636
	Ελάχιστη Τιμή	2.570
	Μέγιστη Τιμή	5.770
Bounce rate	*Μέση τιμή*	25.92%
	Διάμεσος	27.80%
	Τυπική Απόκλιση	7.87%
	Ελάχιστη Τιμή	11.40%
	Μέγιστη Τιμή	52.90%
avg time on site	*Μέση τιμή*	205.258
	Διάμεσος	198.850
	Τυπική Απόκλιση	56.841
	Ελάχιστη Τιμή	95.500
	Μέγιστη Τιμή	369.030
new_visits	*Μέση τιμή*	56.59%
	Διάμεσος	56.55%
	Τυπική Απόκλιση	5.00%
	Ελάχιστη Τιμή	44.36%
	Μέγιστη Τιμή	67.70%
returning visits	*Μέση τιμή*	43.41%
	Διάμεσος	43.45%
	Τυπική Απόκλιση	5.00%
	Ελάχιστη Τιμή	32.30%
	Μέγιστη Τιμή	55.64%
conversions	*Μέση τιμή*	285.246
	Διάμεσος	241.000
	Τυπική Απόκλιση	172.480
	Ελάχιστη Τιμή	22.000
	Μέγιστη Τιμή	802.000
CVR	*Μέση τιμή*	2.73%
	Διάμεσος	2.70%
	Τυπική Απόκλιση	1.03%
	Ελάχιστη Τιμή	0.10%
	Μέγιστη Τιμή	8.00%
conversions SEM	*Μέση τιμή*	139.831
	Διάμεσος	118.000
	Τυπική Απόκλιση	80.486
	Ελάχιστη Τιμή	13.000
	Μέγιστη Τιμή	399.000

Το ποσοστό αναπήδησης (bounce rate) που ορίζει το ποσοστό των χρηστών που βλέπουν μόνο μια σελίδα και στη συνέχεια εγκαταλείπουν (αναπηδούν) το site, χωρίς να έχουν την οποιαδήποτε αλληλεπίδραση με τις λειτουργίες της σελίδας ήταν

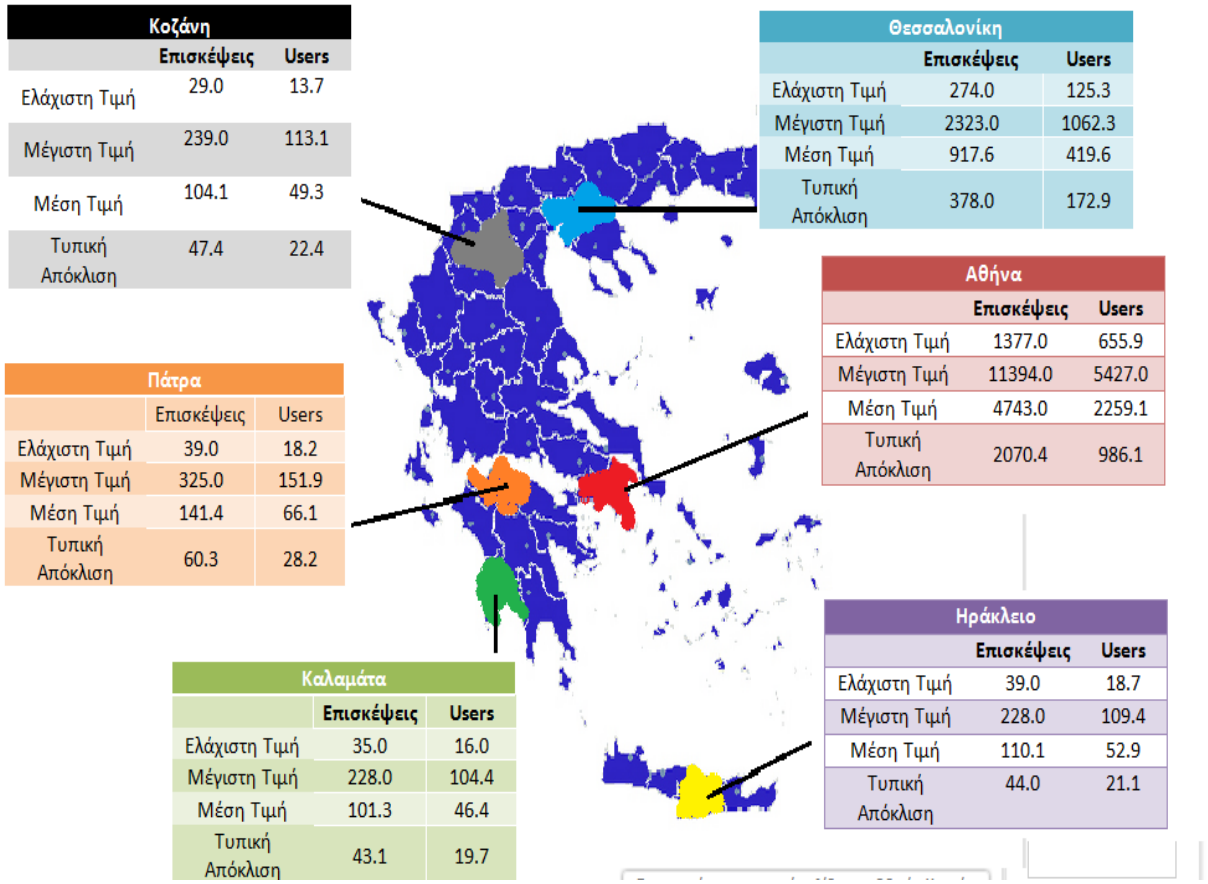
κατά μέσο όρο 25,92% (Τ.Α.=7,87%). Αξίζει να σημειωθεί ότι υπάρχουν ημέρες στο δείγμα μας που σχεδόν οι μισοί επισκέπτες εγκαταλείπουν την ιστοσελίδα, χωρίς να αλληλεπιδράσουν με αυτήν.

Ο μέσος χρόνος επίσκεψης (*avg time on site*) στην ιστοσελίδα είναι 205,258 sec (Τ.Α.= 56,841 sec), το ποσοστό νέων επισκεπτών κατά μέσο όρο προσεγγίζει το 56,69% (Τ.Α.= 5%) και το μέσο ποσοστό των επισκεπτών που έχουν και κατά το παρελθόν επισκεφτεί τον ιστότοπο είναι 43,41%.

Ο μέσος αριθμός στόχων που έχει ο ιστότοπος (μετατροπών) είναι 285,24 (Τ.Α.=172.48) και το μέσο ποσοστό των επισκεπτών που εκτελούν μια μετατροπή (CVR) είναι 2,73% (Τ.Α.=2,70%). Παρατηρείται όμως ότι το ποσοστό CVR παρουσιάζει μεγάλη διακύμανση, αφού η ελάχιστη τιμή του CVR είναι 0,10%, ενώ κάποιες φορές προσέγγισε το 8%.

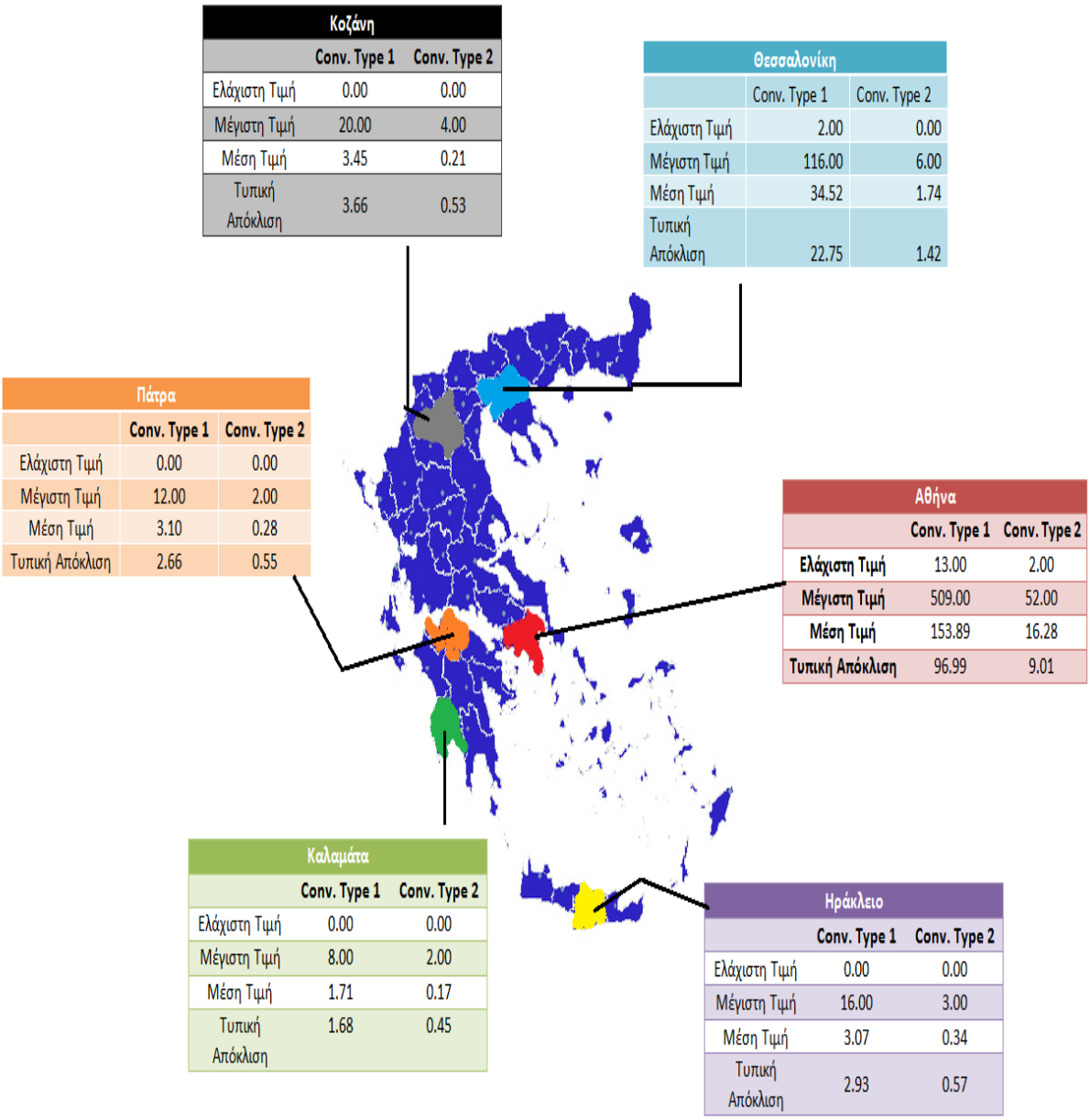
3.2 Διαφοροποιήσεις επισκέψεων στην ιστοσελίδα ανά περιοχή

Στα δεδομένα περιλαμβάνονται δεδομένα από επισκέπτες της ιστοσελίδας που προέρχονται από 6 πόλεις: την Αθήνα, τη Θεσσαλονίκη, την Κοζάνη, την Πάτρα, την Καλαμάτα και το Ηράκλειο. Ο μέσος αριθμός επισκεπτών για το εξεταζόμενο εξάμηνο είναι 7486,35 συνολικά. Από τις εξεταζόμενες πόλεις παρατηρείται ότι η Αθήνα παρουσιάζει μέσο αριθμό επισκεπτών 4743 (Τ.Α.= 2070,4), ακολουθεί η Θεσσαλονίκη με μέσο αριθμό επισκεπτών 917,6, (Τ.Α.= 378) η Πάτρα με 141,4 επισκέπτες (Τ.Α.= 141,4) και το Ηράκλειο με 110,1 επισκέπτες (Τ.Α.=44). Την μικρότερη επισκεψιμότητα παρουσιάζουν οι πόλης της Κοζάνης (Μ.Τ.=104,1, Τ.Α.=47,4) και της Καλαμάτας (Μ.Τ.=101,3, Τ.Α.=43,1).



Εικόνα 1: *Περιγραφικά χαρακτηριστικά για την επισκεψιμότητα και τους χρήστες του ιστότοπου στις έξι πόλεις.*

Όσον αφορά τους χρήστες της ιστοσελίδας διαπιστώνεται ότι η Αθήνα παρουσιάζει πάλι τον μεγαλύτερο αριθμό χρηστών (M.T.=2259,1, T.A.=986,1) και ακολουθούν η Θεσσαλονίκη (M.T.=419,61, T.A.=172,9), η Πάτρα (M.T.=66,1, T.A.=28,2) και το Ηράκλειο (M.T.=52,9, T.A.=21,1). Τους λιγότερους χρήστες έχουν Κοζάνη (M.T.=49,3, T.A.=22,4) και της Καλαμάτας (M.T.=46,4, T.A.=19,7).



Εικόνα 2: *Περιγραφικά χαρακτηριστικά για τις ασφάλειες αυτοκινήτου και υγείας στις έξι πόλεις.*

Ο ιστότοπος που εξετάζουμε, έχει 2 στόχους. Οι στόχοι αυτοί ονομάζονται conversions. Στη συγκεκριμένη περίπτωση το Conversion 1 αποτελεί την Online έκδοση συμβολαίου ασφάλισης αυτοκινήτου, ενώ το Conversion 2 αποτελεί την αρχική εκδήλωση ενδιαφέροντος για την αγορά συμβολαίου ασφάλισης υγείας. Από τα δεδομένα προκύπτει ότι στην Αθήνα κατά μέσο όρο 153,89 online συμβόλαια αυτοκινήτου και στη Θεσσαλονίκη 34,52 online συμβόλαια αυτοκινήτου. Στις υπόλοιπες πόλεις το αντίστοιχο μέγεθος λαμβάνει πολύ χαμηλές τιμές με την

Καλαμάτα να εμφανίζει μέση τιμή ασφαλίσεων αυτοκινήτου 1,71, ενώ για την Κοζάνη, την Πάτρα και το Ηράκλειο οι αντίστοιχες τιμές είναι 3,45, 3,10 και 3,07.

Στην συνέχεια όσον αφορά τα ασφαλιστικά συμβόλαια υγείας κατά μέσο είναι 16,28 για την Αθήνα και 1,74 για τη Θεσσαλονίκη, ενώ οι υπόλοιπες πόλεις παρουσιάζουν αντίστοιχη τιμή μικρότερης της μονάδας. Ο μικρός αριθμός συμβολαίων ασφάλισης υγείας οφείλεται στο ότι η μορφή αυτή συμβολαίων δεν μπορεί να γίνει διαδικτυακά, λόγω διαφόρων περιορισμών.

3.3 Μεταβολές επισκεψιμότητας ανά περιοχή

Από τα αποτελέσματα προκύπτει ότι η μέση μεταβολή της επισκεψιμότητας στον ιστότοπο είναι 3,63% στο Ηράκλειο, 3,04% στην Καλαμάτα, 2,99% στην Κοζάνη, 2,70% στην Πάτρα και 2,69% στην Αθήνα. Τα θετικά ποσοστά που προκύπτουν, δείχνουν την τάση αύξησης της επισκεψιμότητας και στις 6 πόλεις.

Πίνακας 2 *Μεταβολές επισκεψιμότητας ιστότοπου ανά περιοχή*

	Ελάχιστη Τιμή	*Μέγιστη Τιμή*	*Μέση Τιμή*	*Τυπική Απόκλιση*
ΑΘΗΝΑ	-0,72	1,07	0,0269	0,3031
ΘΕΣΣΑΛΟΝΙΚΗ	-0,73	1,26	0,0262	0,2995
ΠΑΤΡΑ	-,073	1,13	0,0270	0,2902
ΚΑΛΑΜΑΤΑ	-0,65	1,21	0,0304	0,2965
ΚΟΖΑΝΗ	-0,77	1,16	0,0299	0,2957
ΗΡΑΚΛΕΙΟ	-0,69	1,43	0,0363	0,3169

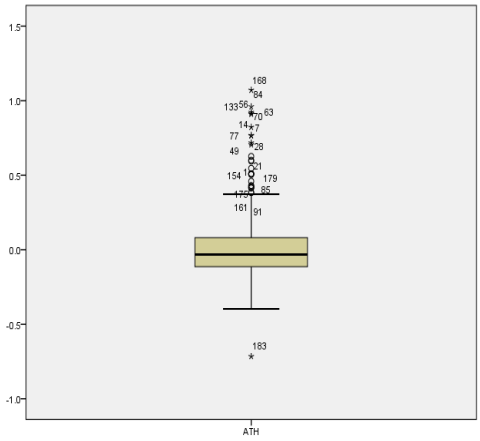
Για να διαπιστώσουμε αν διαφοροποιείται η μεταβολή στους επισκέπτες των 6 πόλεων πραγματοποιήθηκαν έλεγχοι διαφοράς μέσων T-Independent Test, με επίπεδο σημαντικότητας α=10%. Συνολικά πραγματοποιήθηκαν 13 έλεγχοι διαφοράς μέσων, από τους οποίους προέκυψε ότι δεν υπάρχει στατιστικά σημαντική διαφορά στις μεταβολές επισκεψιμότητας στην ιστοσελίδα για τις 6 πόλεις (p-value>10%).

***Πίνακας 3**: Έλεγχος διαφοράς μεταβολών επισκεψιμότητας ιστότοπου ανά περιοχή με T-Independent Samples Test*

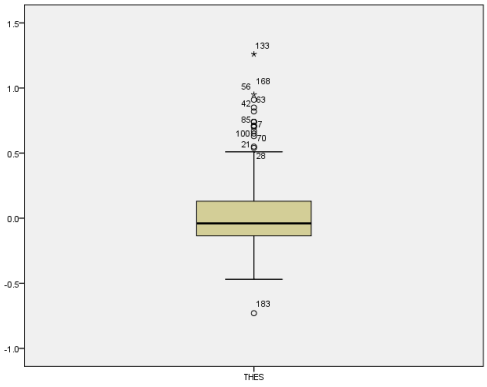
	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
ATKA	.260	186	.795	.00581	-.0383	.0499
ATKO	-.145	186	.885	-.00253	-.0369	.0318
ATHR	.195	184	.846	.01208	-.1102	.1343
THPA	-.039	186	.969	-.00059	-.0306	.0295
THKA	.272	186	.786	.00626	-.0391	.0516
THKO	-.123	186	.902	-.00209	-.0354	.0313
THHR	.205	184	.838	.01276	-.1100	.1355
PAKA	.309	186	.758	.00684	-.0369	.0506
PAKO	-.075	186	.940	-.00150	-.0409	.0379
PAHR	.214	184	.831	.01308	-.1075	.1336
KAKO	-.304	186	.761	-.00834	-.0624	.0457
KAHR	.114	184	.910	.00611	-.0998	.1121
KOHR	.237	184	.813	.01443	-.1055	.1343

Από τα θηκογράμματα (Box-Plots) που ακολουθούν για την κάθε περιοχή ξεχωριστά διαπιστώνεται ότι στην Αθήνα τις περισσότερες ημερομηνίες μέσα στο εξάμηνο υπάρχει μείωση της επισκεψιμότητας. Παρόμοιο φαινόμενο παρατηρείται και για την πόλη της Θεσσαλονίκης. Παρόλο το γεγονός αυτό η μέση τιμή της μεταβολής της επισκεψιμότητας είναι θετική και για τις δύο πόλεις και το αυτό οφείλεται στις πολυάριθμες ακραίες θετικές μεταβολές που εμφανίζει η επισκεψιμότητα της ιστοσελίδας για τις δύο πόλεις.

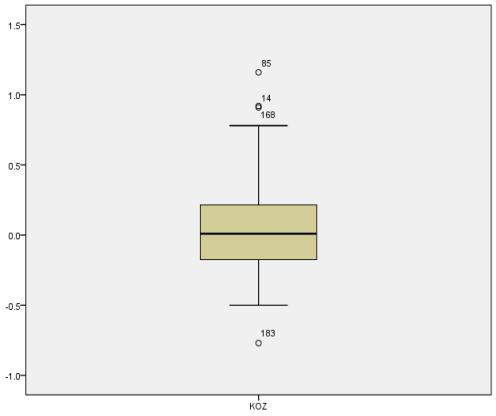
Στις υπόλοιπες τέσσερις πόλεις που συμπεριλαμβάνονται στο δείγμα μας διαπιστώνεται ότι οι μεταβολές των επισκεπτών είναι κατά το ήμισυ θετικές. Επίσης παρουσιάζεται μεγαλύτερη ομαλότητα στα δεδομένα, καθώς σε λίγες περιπτώσεις παρουσιάζονται ακραίες τιμές.



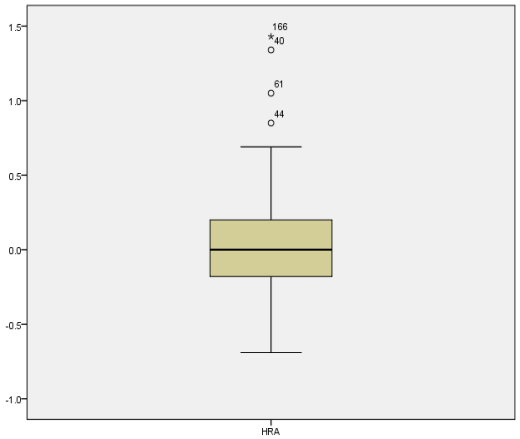
Διάγραμμα 1: Box-Plot για την μεταβολή της επισκεψιμότητας στον ιστότοπο για την περιοχή της Αθήνας



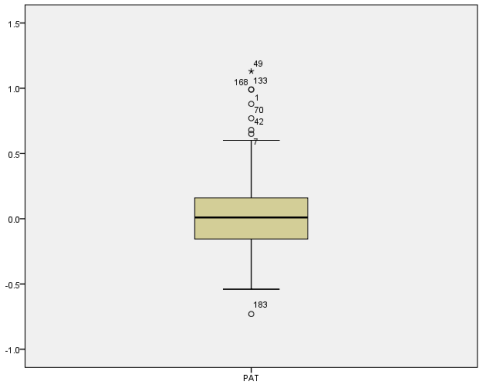
Διάγραμμα 2: Box-Plot για την μεταβολή της επισκεψιμότητας στον ιστότοπο για την περιοχή της Θεσσαλονίκης



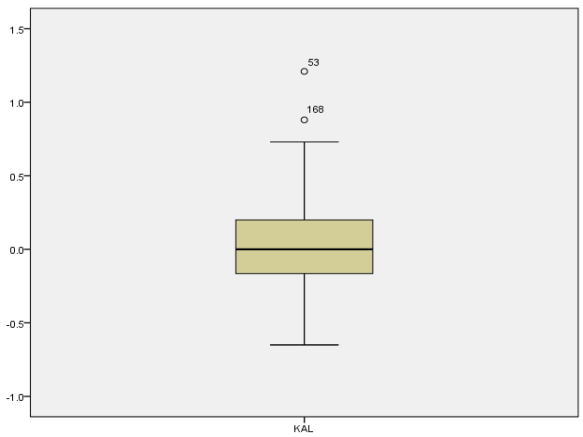
Διάγραμμα 3: Box-Plot για την μεταβολή της επισκεψιμότητας στον ιστότοπο για την περιοχή της Κοζάνης



Διάγραμμα 4 : Box-Plot για την



Διάγραμμα 5: Box-Plot για την μεταβολή της επισκεψιμότητας στον ιστότοπο για την περιοχή της Πάτρας



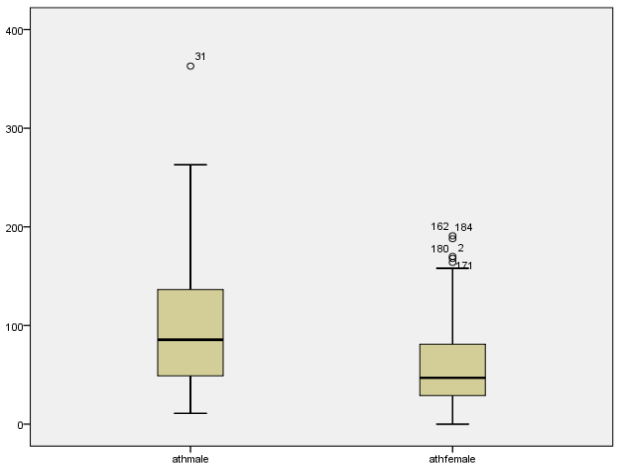
Διάγραμμα 6: Box-Plot για την μεταβολή της επισκεψιμότητας στον ιστότοπο για την περιοχή της Καλαμάτας

3.4 Επίτευξη στόχων – μετατροπών ανά περιοχή

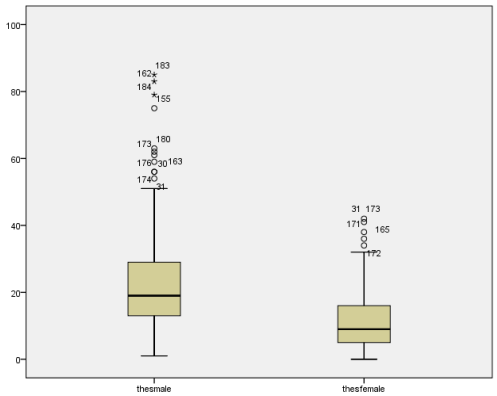
Ακολουθεί η μελέτη της επίδρασης του φύλου του επισκέπτη της ιστοσελίδας στην ανάλυση μας. Συγκεκριμένα, διατίθενται δεδομένα για τον απόλυτο αριθμό των γυναικών και ανδρών που επιλέγουν ασφάλειες αυτοκινήτου και υγείας για κάθε πόλη. Για να διαπιστώσουμε αν διαφοροποιείται ο αριθμός γυναικών και ανδρών που επιλέγουν να κάνουν ασφάλειες αυτοκινήτου, σε κάθε πόλη πραγματοποιούνται έλεγχοι διαφοράς μέσων T-Independent Test, με επίπεδο σημαντικότητας α=10%. Συνολικά πραγματοποιήθηκαν 6 έλεγχοι διαφοράς μέσων (ένας έλεγχος για κάθε πόλη). Τα αποτελέσματα των ελέγχων έδειξαν ότι παρατηρείται στατιστικά σημαντική διαφορά ανάμεσα στα δύο φύλα, σχετικά με την πραγματοποίηση ασφαλίσεων αυτοκινήτου (p-value<10%).

***Πίνακας 4**: Έλεγχος διαφοράς μέσων ανάμεσα στα δύο φύλα ως προς την πραγματοποίηση ασφαλίσεων αυτοκινήτου ανά περιοχή με T-Independent Samples Test*

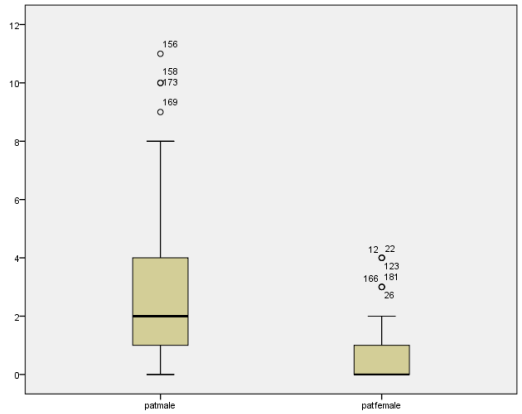
	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
athg1	14.896	183	.000	36.02717	31.2552	40.7991
thesg1	14.287	183	.000	11.41848	9.8416	12.9954
patg1	11.286	183	.000	1.98913	1.6414	2.3369
kalg1	9.415	183	.000	1.04348	.8248	1.2622
kozg1	9.281	183	.000	1.61957	1.2753	1.9639
hrakg1	7.417	183	.000	1.29891	.9534	1.6444



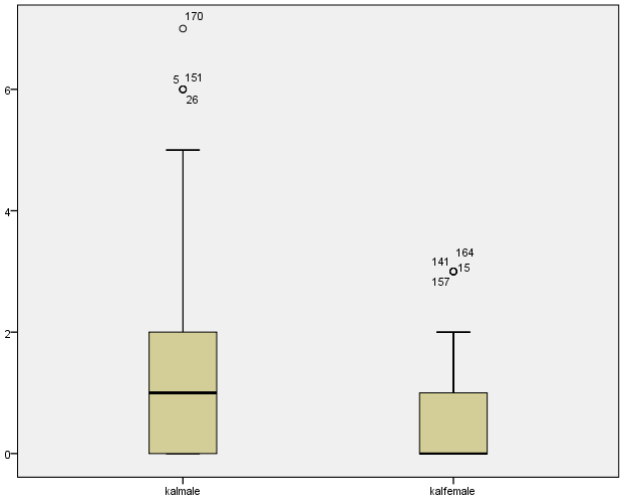
Διάγραμμα 7: Box-Plots για τα δύο φύλα σχετικά με το πλήθος των ασφαλιστηρίων αυτοκινήτων στην περιοχή της Αθήνας



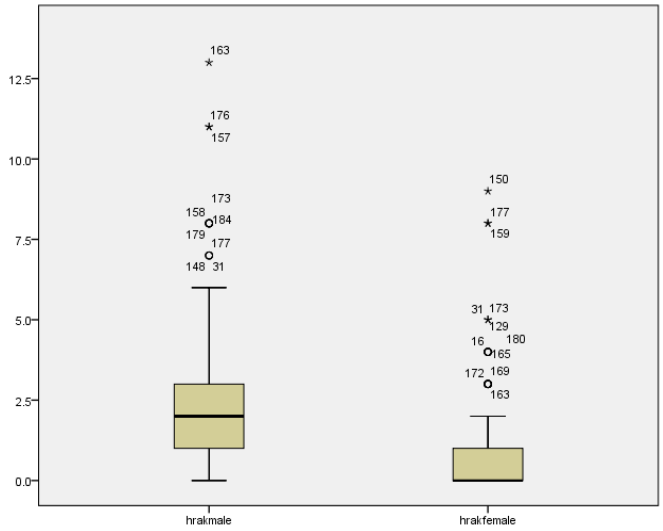
Διάγραμμα 8: Box-Plots για τα δύο φύλα σχετικά με το πλήθος των



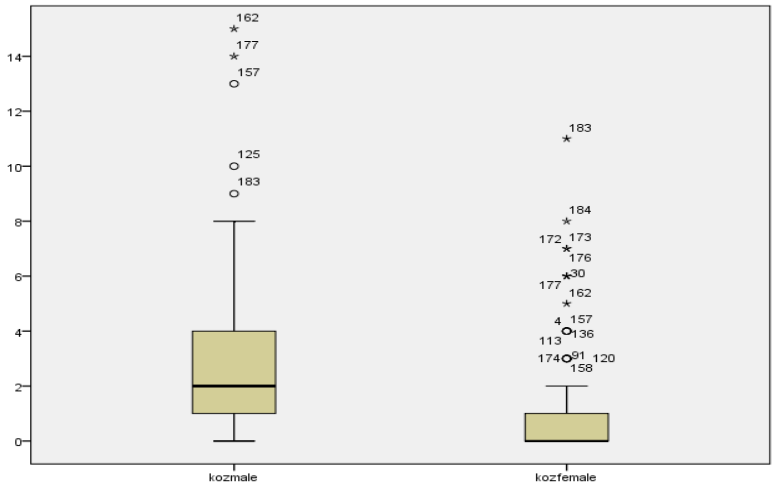
Διάγραμμα 9 Box-Plots για τα δύο φύλα σχετικά με το πλήθος των ασφαλιστηρίων αυτοκινήτων στην περιοχή της Πάτρας



Διάγραμμα 10: Box-Plots για τα δύο φύλα σχετικά με το πλήθος των ασφαλιστηρίων αυτοκινήτων στην περιοχή της Καλαμάτας



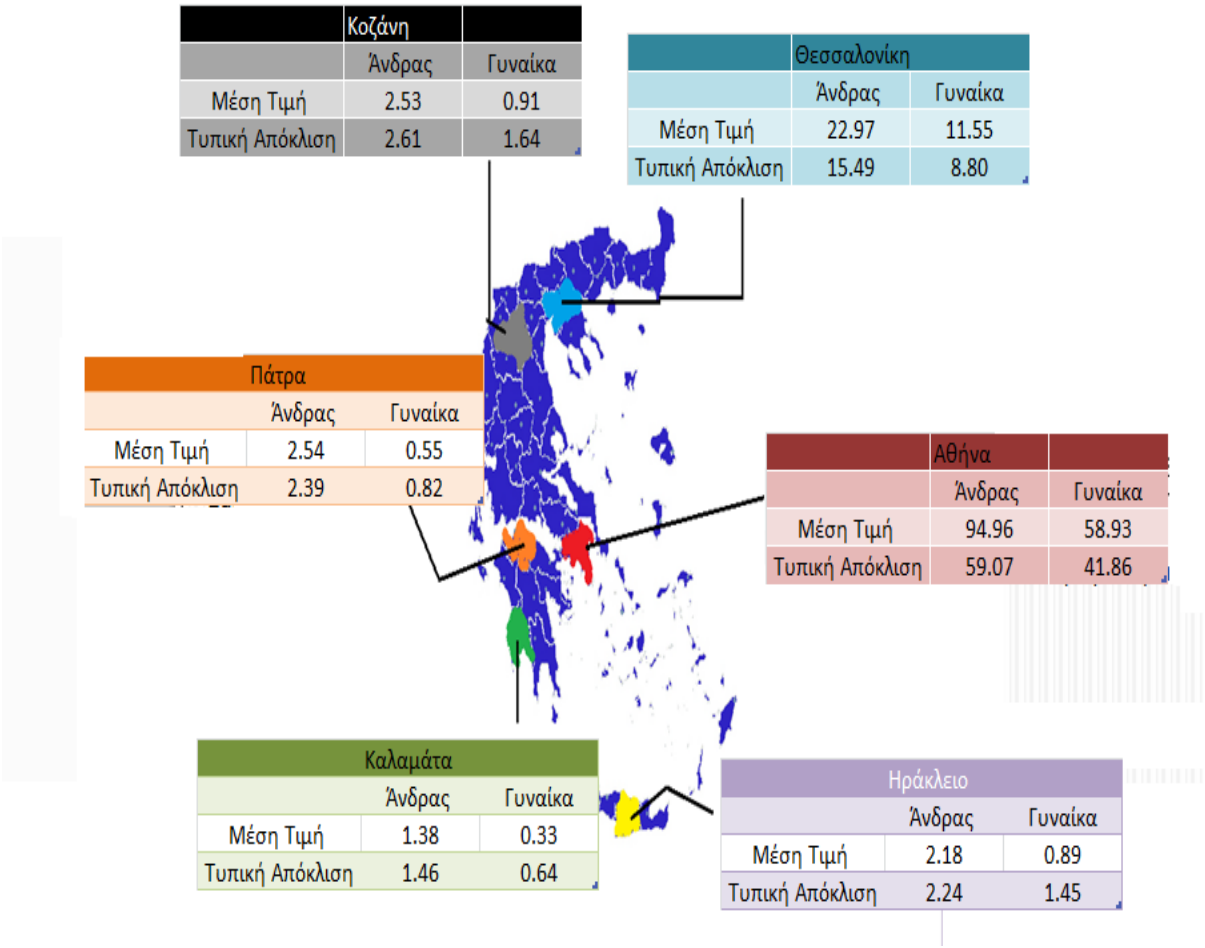
Διάγραμμα 11: Box-Plots για τα δύο φύλα σχετικά το πλήθος των ασφαλιστηρίων αυτοκινήτων στην περιοχή του Ηρακλείου



Διάγραμμα 12: Box-Plots για τα δύο φύλα σχετικά με το πλήθος των ασφαλιστηρίων αυτοκινήτων στην περιοχή της Κοζάνης

Από τα παραπάνω θηκογράμματα διαπιστώνεται ότι οι άνδρες παρουσιάζουν μεγαλύτερη τάση από τις γυναίκες να πραγματοποιούν ασφαλιστήρια αυτοκινήτου διαδικτυακά. Στην Καλαμάτα, στην Πάτρα, στην Κοζάνη και στο Ηράκλειο διαπιστώνεται ότι στις μισές παρατηρήσεις οι γυναίκες παρουσιάζουν την ελάχιστη τιμή μετατροπών (η διάμεσος συμπίπτει με τη ελάχιστη τιμή). Η Θεσσαλονίκη είναι η

μόνη πόλη που οι άνδρες εμφανίζουν ακραίες υψηλές τιμές, ενώ οι γυναίκες παρουσιάζουν παρόμοιο φαινόμενο στην Κοζάνη και στο Ηράκλειο.



Εικόνα 3: *Περιγραφικά χαρακτηριστικά για τις ασφάλειες αυτοκινήτων για γυναίκες και άνδρες στις έξι πόλεις.*

Από τον παραπάνω πίνακα προκύπτει ότι οι άνδρες εμφανίζουν μεγαλύτερη τάση να πραγματοποιούν ασφάλειες αυτοκινήτου σε σχέση με τις γυναίκες.

Για να διαπιστώσουμε αν διαφοροποιείται ο αριθμός γυναικών και ανδρών που επιλέγουν να κάνουν ασφάλειες υγείας διαδικτυακά, σε κάθε πόλη πραγματοποιούνται έλεγχοι διαφοράς μέσων T-Independent Test, με επίπεδο σημαντικότητας α=10%. Συνολικά πραγματοποιήθηκαν 6 έλεγχοι διαφοράς μέσων (ένας έλεγχος για κάθε πόλη). Τα αποτελέσματα των ελέγχων έδειξαν ότι δεν παρατηρείται στατιστικά σημαντική διαφορά ανάμεσα στα δύο φύλα (p-value>10%).

3.5 Πρόβλεψη του CVR με την εφαρμογή παλινδρόμησης

Στο κεφάλαιο αυτό πραγματοποιείται ανάλυση παλινδρόμησης, ώστε να προσδιοριστεί ένα υπόδειγμα, στο οποίο θα εκτιμάται το CVR, με ανεξάρτητες μεταβλητές τους επισκέπτες (*visitors*), τους μοναδικούς επισκέπτες (*unique visitors*), τον συνολικό αριθμό σελίδων που επισκέφτηκαν οι επισκέπτες (*pageviews***)**, ο μέσος αριθμός σελίδων που επισκέπτεται ένας επισκέπτης (*pages/visit*), το ποσοστό αναπήδησης που ορίζει το ποσοστό των χρηστών που βλέπουν μόνο μια σελίδα και στη συνέχεια εγκαταλείπουν το site (*bounce rate*), τον μέσο χρόνο επίσκεψης (*avg. time on site*), το ποσοστό νέων επισκέψεων (*new visits*) και το ποσοστό επισκεπτών που έχουν και κατά το παρελθόν επισκεφτεί τον ιστότοπο (*returning visits*).

***Πίνακας 6**: Πίνακας Συσχετίσεων μεταξύ των μεταβλητών (ανεξάρτητων και εξαρτημένης.*

Correlations

		CVR	visitors	unique_visitors	pageviews	pages_per_visit	bounce_rate	avg_time_on_site	new_visits	returning_visits
Pearson Correlation	CVR	1.000	-.144	-.177	.022	.352	-.010	.577	-.228	.018
	visitors	-.144	1.000	.999	.898	-.048	-.125	-.094	.992	.972
	unique_visitors	-.177	.999	1.000	.887	-.071	-.120	-.125	.996	.960
	pageviews	.022	.898	.887	1.000	.379	-.432	.314	.864	.923
	pages_per_visit	.352	-.048	-.071	.379	1.000	-.701	.946	-.105	.062
	bounce_rate	-.010	-.125	-.120	-.432	-.701	1.000	-.541	-.100	-.167
	avg_time_on_site	.577	-.094	-.125	.314	.946	-.541	1.000	-.169	.051
	new_visits	-.228	.992	.996	.864	-.105	-.100	-.169	1.000	.934
	returning_visits	.018	.972	.960	.923	.062	-.167	.051	.934	1.000
Sig. (1-tailed)	CVR	.	.026	.008	.381	.000	.445	.000	.001	.406
	visitors	.026	.	.000	.000	.261	.046	.104	.000	.000
	unique_visitors	.008	.000	.	.000	.171	.052	.046	.000	.000
	pageviews	.381	.000	.000	.	.000	.000	.000	.000	.000
	pages_per_visit	.000	.261	.171	.000	.	.000	.000	.079	.204
	bounce_rate	.445	.046	.052	.000	.000	.	.000	.089	.012
	avg_time_on_site	.000	.104	.046	.000	.000	.000	.	.011	.247
	new_visits	.001	.000	.000	.000	.079	.089	.011	.	.000
	returning_visits	.406	.000	.000	.000	.204	.012	.247	.000	.
N	CVR	183	183	183	183	183	183	183	183	183
	visitors	183	183	183	183	183	183	183	183	183
	unique_visitors	183	183	183	183	183	183	183	183	183
	pageviews	183	183	183	183	183	183	183	183	183
	pages_per_visit	183	183	183	183	183	183	183	183	183
	bounce_rate	183	183	183	183	183	183	183	183	183
	avg_time_on_site	183	183	183	183	183	183	183	183	183
	new_visits	183	183	183	183	183	183	183	183	183
	returning_visits	183	183	183	183	183	183	183	183	183

Για να είναι αξιόπιστο το υπόδειγμα που θα εκτιμηθεί θα πρέπει να ελεγχθεί αν επαληθεύονται οι υποθέσεις της πολλαπλής παλινδρόμησης. Από τον παραπάνω πίνακα παρατηρείται ότι μεταξύ των μεταβλητών δεν παρατηρείται σημαντική

συσχέτιση, παρά μόνο μεταξύ των μεταβλητών «visitors» και «unique visitors», όπου δεν προκαλεί πρόβλημα στο υπόδειγμα μας, καθώς η μεταβλητή «visitors» δεν περιλαμβάνεται στο τελικό μας υπόδειγμα. Επομένως, δεν υπάρχει πολυσυγγραμμικότητα στο υπόδειγμα.

Στο υπόδειγμα δεν υπάρχει αυτοσυσχέτιση, καθώς ο συντελεστής Durbin-Watson είναι κοντά στο 2.

Πίνακας 7: Model Summary

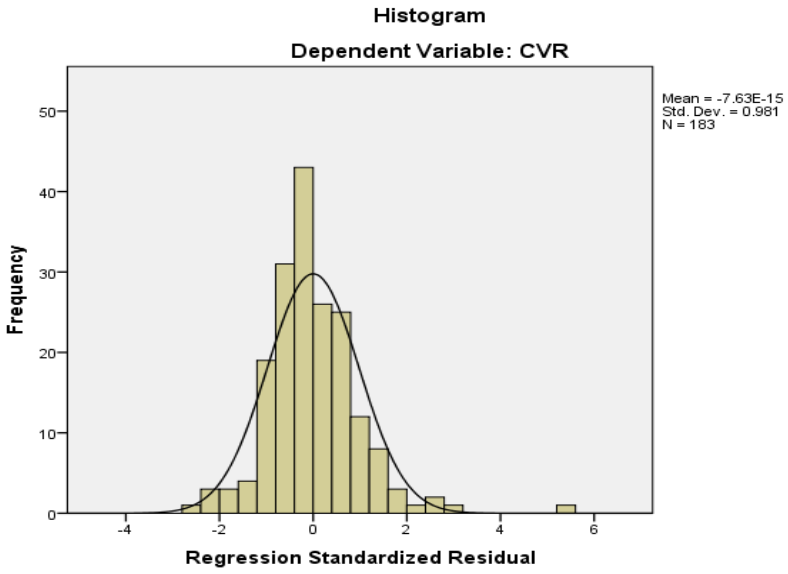
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.849[a]	.721	.710	.006	1.814

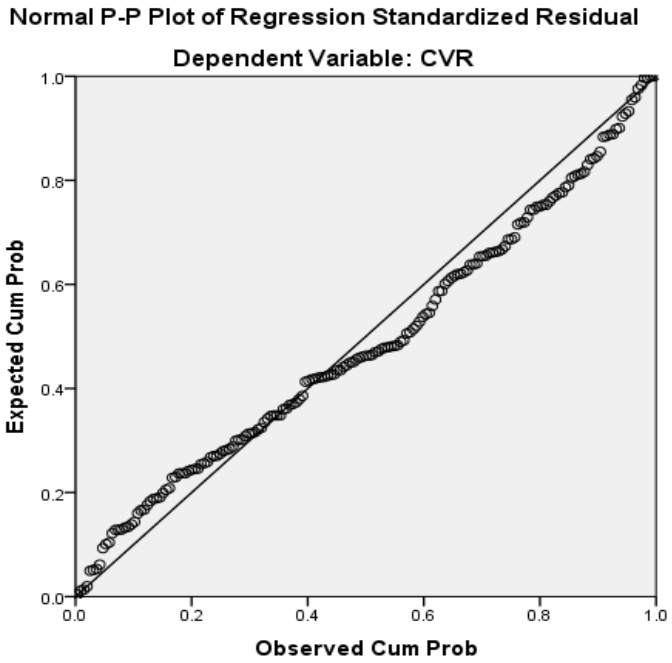
a. Predictors: (Constant), returning_visits, avg_time_on_site, bounce_rate, new_visits, pages_per_visit, pageviews, unique_visitors

b. Dependent Variable: CVR

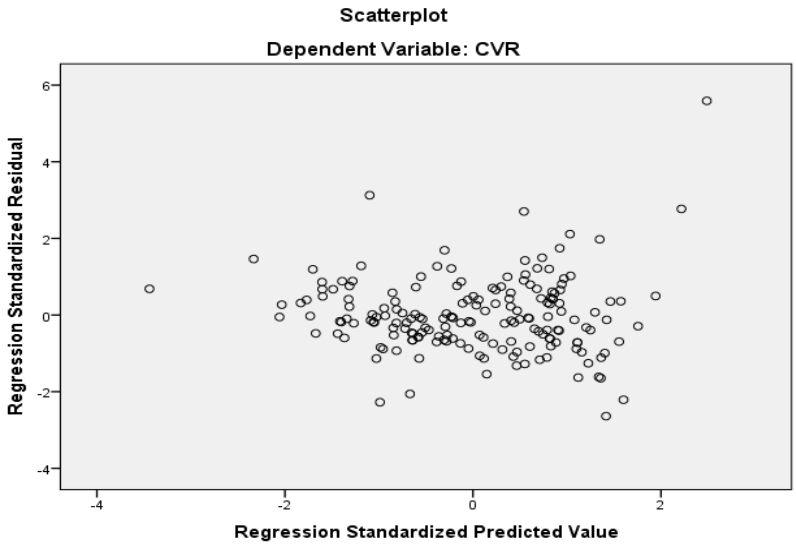
Επιπλέον, τα κατάλοιπα ακολουθούν την κανονική κατανομή όπως προκύπτει από το ιστόγραμμα και από το P-P Plot.

Διάγραμμα 13: Ιστόγραμμα Καταλοίπων.





Διάγραμμα 14: P-P Plot Καταλοίπων.



Διάγραμμα 15: *Διάγραμμα Διασποράς*

Η οπτική επιβεβαίωση της τυχαίας κίνησης των καταλοίπων γίνεται μέσω του διαγράμματος σημείων (scatter plot) της εξαρτημένης μεταβλητής με τις τιμές των λαθών (ή καταλοίπων). Η εικόνα του διαγράμματος σημείων είναι τυχαία και δεν περιγράφει κάποια συστηματική κίνηση. Άρα δεν υπάρχει κάποιος συστηματικός

παράγοντας που επιδρά στην κίνηση της εξαρτημένης μεταβλητής που δεν συμμετάσχει στο μοντέλο της παλινδρόμησης.

Από τον πίνακα ANOVA το υπόδειγμα είναι στατιστικά σημαντικό (F=65,641, p-value <1%) και επομένως έχει νόημα να χρησιμοποιηθεί για την πρόβλεψη του CVR.

Πίνακας 8: *Πίνακας* ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.014	7	.002	64.541	.000[b]
	Residual	.005	175	.000		
	Total	.019	182			

a. Dependent Variable: CVR

b. Predictors: (Constant), returning_visits, avg_time_on_site, bounce_rate, new_visits, pages_per_visit, pageviews, unique_visitors

Στον πίνακα *Model Summary* απεικονίζονται μερικοί βασικοί δείκτες καλής προσαρμογής του μοντέλου. Ο δείκτης R Square είναι ένδειξη του ποσοστού της διακύμανσης της εξαρτημένης μεταβλητής που επεξηγεί το μοντέλο. Στο τελευταίο στάδιο κατασκευής του το μοντέλο της γραμμικής παλινδρόμησης επεξηγεί το 72,1% της συνολικής διακύμανσης του δείγματος. Τιμές του δείκτη κοντά στο ένα είναι ένδειξη ότι οι παράγοντες που συμμετέχουν στην διαδικασία κατασκευής του μοντέλου είναι ικανοποιητικοί για την περιγραφή της κίνησης της εξαρτημένης μεταβλητής.

Το υπόδειγμα πρόβλεψης του CVR είναι το εξής:

$$CVR = 0{,}049000 + 0{,}000005 * unique\ visitors - 0{,}0000001 * pageviews - 0{,}028 * pages\ per\ visit - 0{,}012 * bounce\ rate + 0{,}002 * avg\ time\ on\ site - 0{,}000008 * new\ visits + 0{,}000002 * returning\ visits$$

Πίνακας 9: *Πίνακας Παραμέτρων*

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.049	.010		5.018	.000
	unique_visitors	5.928E-006	.000	1.686	.988	.324
	pageviews	-1.470E-007	.000	-.164	-.470	.639
	pages_per_visit	-.028	.004	-1.714	-7.647	.000
	bounce_rate	-.012	.009	-.092	-1.280	.202
	avg_time_on_site	.002	.000	2.092	11.407	.000
	new_visits	-8.746E-006	.000	-1.817	-1.410	.160
	returning_visits	2.074E-006	.000	.230	.473	.637

a. Dependent Variable: CVR

Από τα παραπάνω προκύπτει ότι η επίδραση της εκάστοτε ανεξάρτητης μεταβλητής, στο CVR όταν οι άλλοι παράγοντες παραμείνουν σταθεροί είναι η εξής:

- Όταν οι μοναδικοί επισκέπτες αυξηθούν κατά μια μονάδα, τότε το CVR αυξάνεται κατά 0,000005.
- Όταν ο συνολικός αριθμός σελίδων που επισκέφτηκαν οι επισκέπτες αυξηθεί κατά μια μονάδα, τότε το CVR μειώνεται κατά 0,0000001.
- Όταν ο μέσος όρος σελίδων που επισκέπτεται ένας επισκέπτης αυξηθεί κατά μια μονάδα, τότε το CVR μειώνεται κατά 0,028.
- Όταν το ποσοστό των χρηστών που βλέπουν μόνο μια σελίδα και στη συνέχεια εγκαταλείπουν αυξηθεί κατά μια μονάδα, τότε το CVR μειώνεται κατά 0,012.
- Όταν ο μέσος χρόνος επίσκεψης αυξηθεί κατά μια μονάδα, τότε το CVR αυξάνεται κατά 0,002.
- Όταν το ποσοστό νέων επισκεπτών αυξηθεί κατά μία μονάδα, τότε το CVR μειώνεται κατά 0,000008.
- Όταν το ποσοστό επισκεπτών που έχουν και κατά το παρελθόν επισκεφτεί τον ιστότοπο αυξηθεί κατά μία μονάδα, τότε το CVR αυξάνεται κατά 0,000002.

3.6 Συσταδοποίηση

Στόχος του εμπειρικού μέρους είναι ο προσδιορισμός των υπο-ομάδων, με κριτήριο την επισκεψιμότητα της ιστοσελίδας μέσα στην εξεταζόμενη περίοδο. Τα δεδομένα που χρησιμοποιήσαμε για τον προσδιορισμό των συστάδων, είναι το πλήθος των επισκεπτών ανά ημέρα, των «μοναδικών» επισκεπτών, το συνολικό αριθμό σελίδων, το μέσο χρόνο που μένει κάποιος στην ιστοσελίδα όταν την επισκέπτεται, τους νέους επισκέπτες, το ποσοστό αναπήδησης (bounce rate), το μέσο ποσοστό των επισκεπτών που έχουν και κατά το παρελθόν επισκεφτεί τον ιστότοπο, το μέσο αριθμό στόχων που έχει ο ιστότοπος (μετατροπών) και το μέσο ποσοστό των επισκεπτών που εκτελούν μια μετατροπή (CVR).

Παρακάτω παρατίθενται τα αποτελέσματα της συσταδοποίησης όπως προέκυψαν με την εφαρμογή του αλγόριθμου «K-Means», όπου ο αριθμός των συστάδων που ορίστηκε στον παραπάνω αλγόριθμο για κάθε σενάριο είχε αρχικά προσδιοριστεί με την ιεραρχική μέθοδο.

Η συσταδοποίηση έγινε με βάση διαφορετικά σενάρια:

- 1^{ο} Σενάριο: Συσταδοποίηση ημερών με κριτήριο το πλήθων των επισκεπτών/ημέρα
- 2^{ο} Σενάριο: Συσταδοποίηση ημερών με κριτήριο το πλήθων των νέων επισκεπτών/ημέρα
- 3^{ο} Σενάριο: Συσταδοποίηση ημερών με κριτήριο το μέσο αριθμό στόχων που έχει ο ιστότοπος (μετατροπών).
- 4^{ο} Σενάριο: Συσταδοποίηση ημερών με κριτήριο το πλήθος των επισκεπτών ανά ημέρα, των «μοναδικών» επισκεπτών, τον συνολικό αριθμό σελίδων, το μέσο χρόνο που μένει κάποιος στην ιστοσελίδα όταν την επισκέπτεται, τους νέους επισκέπτες, το ποσοστό αναπήδησης (bounce rate), το μέσο ποσοστό των επισκεπτών που έχουν και κατά το παρελθόν επισκεφτεί τον ιστότοπο, το μέσο αριθμό στόχων που έχει ο ιστότοπος (μετατροπών) και το μέσο ποσοστό των επισκεπτών που εκτελούν μια μετατροπή (CVR).

3.6.1 Σενάριο 1: Συσταδοποίηση ημερών με κριτήριο το πλήθων των επισκεπτών/ημέρα

Συγκεκριμένα, χρησιμοποιώντας το SPSS πραγματοποιήθηκε Ανάλυση Cluster και συγκεκριμένα ο αλγόριθμος k-means. Με τον τρόπο αυτό επιδιώκεται να πραγματοποιηθεί μια καθαρά data mining τεχνική σε αντιδιαστολή με τις στατιστικές τεχνικές που έχουν χρησιμοποιηθεί παραπάνω.

Συγκεκριμένα, επιδιώξαμε να διακρίνουμε τις υπο-ομάδες την χρονική περίοδο για την οποία έχουμε αντλήσει τα δεδομένα, έτσι ώστε να μελετήσουμε πως διαφοροποιείται η επισκεψιμότητα της ιστοσελίδας μέσα στην εξεταζόμενη περίοδο. Τα δεδομένα που χρησιμοποιήσαμε για τον προσδιορισμό των συστάδων, είναι το πλήθος των επισκεπτών ανά ημέρα. Με τον τρόπο αυτό επιδιώκεται να αναζητηθεί τρόπος ομαδοποίησης των χρονικών περιόδων με κριτήριο την επισκεψιμότητα.

Από την ανάλυση προέκυψαν τρεις ομάδες, στην πρώτη ομάδα αντιστοιχούν οι μέρες με χαμηλή επισκεψιμότητα, στη δεύτερη ανήκουν οι χρονικές περίοδοι που η επισκεψιμότητα είναι μέτρια και στην τρίτη ομάδα η περίοδος όπου η επισκεψιμότητα είναι έντονη. Στην πρώτη ομάδα ανήκει η περίοδος Αύγουστος-Σεπτέμβριος. Την περίοδο αυτή ο μέσος αριθμός επισκεπτών για το συγκεκριμένο εξάμηνο ήταν 4955,14 επισκέπτες. Στη δεύτερη ομάδα ανήκουν οι μήνες Ιούλιος και Οκτώβριος με μέση επισκεψιμότητα 7100,12 επισκέπτες και στην τρίτη ομάδα ανήκει η χρονική περίοδο Νοέμβριος-Δεκέμβριος 2014 με μέση επισκεψιμότητα 11.234,32 επισκέπτες.

3.6.2 Σενάριο 2: Συσταδοποίηση ημερών με κριτήριο το πλήθων των νέων επισκεπτών/ημέρα

Στην περίπτωση αυτή επιδιώξαμε να διακρίνουμε τις υπο-ομάδες την χρονική περίοδο για την οποία έχουμε αντλήσει τα δεδομένα, έτσι ώστε να μελετήσουμε πως διαφοροποιείται η επισκεψιμότητα της ιστοσελίδας μέσα στην εξεταζόμενη περίοδο με κριτήριο τους νέους επισκέπτες.

Από την ανάλυση προέκυψαν δύο ομάδες, στην πρώτη ομάδα αντιστοιχούν οι μήνες: Ιούλιος, Αύγουστος, Σεπτέμβριος με μέση τιμή τιμή 4.765,32 νέοι επισκέπτες

και στην δεύτερη ομάδα η χρονική περίοδος Οκτώβριος-Δεκέμβριος 2014 με τους νέους επισκέπτες να προσεγγίζουν τους 8.435,67 νέοι επισκέπτες/ημέρα.

3.6.3 Σενάριο 3: Συσταδοποίηση ημερών με κριτήριο το μέσο αριθμό στόχων που έχει ο ιστότοπος (μετατροπών)

Στην συνέχεια επιδιώκεται, να διακρίνουμε τις υπο-ομάδες την χρονική περίοδο για την οποία έχουμε αντλήσει τα δεδομένα, έτσι ώστε να μελετήσουμε πως διαφοροποιείται ο μέσος αριθμός στόχων (μετατροπών). Με τον τρόπο αυτό επιδιώκεται να αναζητηθεί ο τρόπος ομαδοποίησης των χρονικών περιόδων με το παραπάνω κριτήριο.

Από την ανάλυση προέκυψαν τρεις ομάδες, στην πρώτη ομάδα αντιστοιχούν οι μέρες πολύ χαμηλές μετατροπές, και η περίοδος που αντιστοιχεί στην ομάδα αυτή είναι ο Αύγουστος 2014 με μέσο αριθμό μετατροπών 162,45 και στη δεύτερη ανήκει η περίοδος που περιλαμβάνει τον Ιούλιο, τον Σεπτέμβριο-Οκτώβριο με μέσο αριθμό μετατροπών 234,12 και η τρίτη ομάδα περιλαμβάνει το Δεκέμβριο και το μέσο πλήθος μετατροπών είναι 456,78.

3.6.4 Σενάριο 4: Πολυκριτηριακή Συσταδοποίηση ημερών

Στο σενάριο αυτό εξετάζεται πως μπορούν οι μέρες να ταξινομηθούν με ένα σύνολο πληροφοριών που έχουν συλλεχθεί και αφορούν την ιστοσελίδα, όπως είναι το πλήθος των επισκεπτών ανά ημέρα, τον «μοναδικών» επισκεπτών, τον συνολικό αριθμό σελίδων, το μέσο χρόνο που μένει κάποιος στην ιστοσελίδα όταν την επισκέπτεται, τους νέους επισκέπτες, το ποσοστό αναπήδησης (bounce rate), το μέσο ποσοστό των επισκεπτών που έχουν και κατά το παρελθόν επισκεφτεί τον ιστότοπο, το μέσο αριθμό στόχων που έχει ο ιστότοπος (μετατροπών) και το μέσο ποσοστό των επισκεπτών που εκτελούν μια μετατροπή (CVR).

- **Αλγόριθμος k-means**

Συσταδοποίηση οι χρονικές περίοδοι που η επισκεψιμότητα είναι μέτρια και στην τρίτη ομάδα η περίοδος όπου η επισκεψιμότητα είναι έντονη. Στην πρώτη ομάδα ανήκει η περίοδος Αύγουστος-Σεπτέμβριος. Την περίοδο αυτή ο μέσος

αριθμός επισκεπτών για το συγκεκριμένο εξάμηνο ήταν 4955,14 επισκέπτες. Στη δεύτερη ομάδα ανήκουν οι μήνες Ιούλιος και Σεπτέμβριος με μέση επισκεψιμότητα 7100,12 επισκέπτες και στην τρίτη ομάδα ανήκει η χρονική περίοδο Νοέμβριος-Δεκέμβριος 2014 με μέση επισκεψιμότητα 11.234,32 επισκέπτες.

Οι μοναδικοί επισκέπτες που μετρήθηκαν με βάση την μοναδική IP, η οποία είναι η ηλεκτρονική ταυτότητα του κάθε επισκέπτη είναι κατά μέσο όρο 4245,14 επισκέπτες και ο συνολικός αριθμός σελίδων που επισκέφτηκαν οι επισκέπτες είναι κατά μέσο όρο 16703,16. Ο μέσος όρος σελίδων που επισκέπτεται ένας επισκέπτης είναι 3,38. Το ποσοστό αναπήδησης (bounce rate) ήταν κατά μέσο όρο 28,84%. Ο μέσος χρόνος επίσκεψης (*avg time on site*) στην ιστοσελίδα είναι 188,4 sec και ο μέσος όρος στόχων που έχει ο ιστότοπος (μετατροπών) είναι 157,37.

Final Cluster Centers

	Cluster	
	1	2
visitors	4955.14	9982.42
unique_visitors	4245.14	8721.40
pageviews	16703.16	36680.11
pages_per_visit	3.38	3.80
bounce_rate	.2884	.2308
avg_time_on_site	188.40	221.55
new_visits	2767.67	5912.46
returning_visits	2187.47	4069.97
conversions	157.37	410.22

Στην δεύτερη ομάδα ανήκουν οι ημερομηνίες στις οποίες, το κοινό δείχνει μεγάλη τάση να επισκέπτεται την ιστοσελίδα. Την περίοδο αυτή τα μεγέθη που εξετάζονται παρουσιάζουν διπλάσιες τιμές.

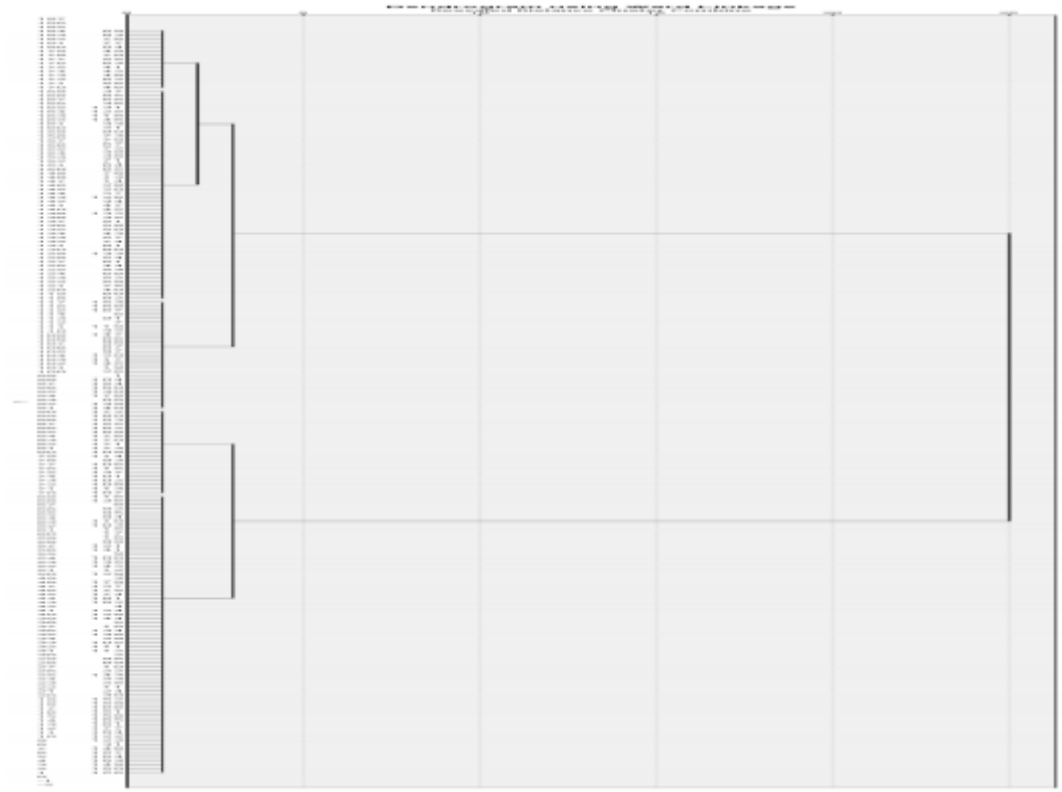
- **Ιεραρχική Μέθοδος**

Στην συνέχεια θα πραγματοποιηθεί ανάλυση συστάδων με την εφαρμογή της ιεραρχικής μεθόδου. Στην εφαρμογή της μεθόδου αυτής αρχικά επιλέγεται ο αριθμός των συστάδων. Από τον παρακάτω πίνακα διαπιστώνουμε ότι στο δεύτερο στάδιο παρατηρείται μια απότομη αύξηση στην τιμή του «Coefficient», επομένως

επιλέγουμε δύο συστάδες, για να ομαδοποιήσουμε τις ημέρες που περιλαμβάνονται στο δείγμα.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	69	83	2037.950	0	0	11
2	51	58	4897.014	0	0	7
3	49	70	8357.298	0	0	13
4	74	81	15890.861	0	0	18
5	64	65	26564.514	0	0	35
6	67	72	37981.962	0	0	43
7	50	51	50612.669	0	2	32
8	55	63	64328.259	0	0	13
9	43	57	79086.972	0	0	32
10	52	59	96543.345	0	0	19
11	69	76	116117.745	1	0	63
12	48	62	136099.901	0	0	45



Αφού πραγματοποιηθεί ομαδοποίηση με τη μέθοδο που περιγράψαμε παραπάνω, γίνεται περιγραφική ανάλυση των συστάδων που προκύπτουν.

Συγκεκριμένα, οι μέρες που συμπεριλάβαμε στο δείγμα μας χωρίζονται σε δυο κατηγορίες: σε αυτές που το πλήθος των επισκεπτών (M=10017 vs. M=5028), των «μοναδικών επισκεπτών» (M=8753 vs. M=4309), τον συνολικό αριθμό σελίδων (M=36862 vs. M=16953), το μέσο χρόνο που μένει κάποιος στην ιστοσελίδα όταν την επισκέπτεται (M=3,80 vs. M=3,37), τους νέους επισκέπτες (M=5941 vs. M=2806), το ποσοστό αναπήδησης (bounce rate) (M=0.25 vs. M=0.22), το μέσο ποσοστό των επισκεπτών που έχουν και κατά το παρελθόν επισκεφτεί τον ιστότοπο (M=4075 vs. M=2221) και το μέσο ποσοστό των επισκεπτών που εκτελούν μια μετατροπή (CVR) (M=410 vs. M=162) εμφανίζουν υψηλότερες τιμές στην μια ομάδα γεγονός που δείχνει ότι μιλάμε για περιόδους υψηλής επισκεψιμότητας, σε σχέση μάλιστα με τις περιόδους που ανήκουν στην δεύτερη ομάδα.

		N	Mean	Std. Deviation
visitors	1	94	5028.9468	1704.58954
	2	90	10017.0556	2361.49597
	Total	184	7468.7826	3231.11734
unique_visitors	1	94	4309.7234	1572.64559
	2	90	8753.4222	2198.46424
	Total	184	6483.2717	2927.23743
pageviews	1	94	16953.6702	5721.79880
	2	90	36862.4000	5522.16086
	Total	184	26691.6359	11447.80123
pages_per_visit	1	94	3.3754	.39061
	2	90	3.8069	.75598
	Total	184	3.5865	.63424
bounce_rate	1	94	.288862	.0614680
	2	90	.228978	.0831678
	Total	184	.259571	.0786452
avg_time_on_site	1	94	188.7315	40.93889
	2	90	221.9447	65.69423
	Total	184	204.9771	56.81370
new_visits	1	94	2806.9681	1165.72585
	2	90	5941.3000	1701.54458
	Total	184	4340.0652	2137.03868

	1	94	2221.9787	602.31078
returning_visits	2	90	4075.7556	731.02905
	Total	184	3128.7174	1143.53506
	1	94	162.3404	73.47307
conversions	2	90	410.6444	155.37503
	Total	184	283.7935	173.13266

Συμπεράσματα – Προτάσεις

Ποιοι είναι άραγε οι λόγοι για τους οποίους θα θέλαμε να κάνουμε εξόρυξη γνώσης στα δεδομένα χρήσης του παγκόσμιου ιστού, κυρίως για την εξατομίκευση (personalization) της διαδικτυακής εμπειρίας του χρήστη ενός ιστοτόπου. Δηλαδή σε κάθε χρήστη να εμφανίζεται διαφορετικά η ιστοσελίδα ανάλογα με το πώς αυτός κινείται μεταξύ των συνδέσμων της. Πιθανοί λόγοι για την εξατομίκευση μπορεί να είναι, ανάλογα και με το σκοπό μας:

- Να τον διευκολύνουμε προτείνοντάς του συνδέσμους σε άλλες σελίδες που τον ενδιαφέρουν χωρίς να αναγκαστεί να τις ψάξει. Π.χ. με ένα πλαίσιο στη δεξιά ή αριστερή πλευρά της σελίδας που θα ονομάζεται "προτεινόμενες σελίδες" και θα του δείχνει σελίδες που πιθανώς τον ενδιαφέρουν.

- Να του προτείνουμε με τρόπο παρόμοιο με της προηγούμενης περίπτωσης σελίδες που ξέρουμε ότι πιθανώς δεν τον ενδιαφέρουν σε μια προσπάθειά μας να αυξήσουμε και την επισκεψιμότητα άλλων σελίδων.

- Να εμφανίζονται διαφημίσεις στη σελίδα όχι απαραίτητα σχετικές με την τρέχουσα αλλά και με άλλες σελίδες που πιθανολογούμε ότι θα επισκεφθεί ο χρήστης μας.

- Βελτίωση της λειτουργικότητας του ιστοτόπου. Από την ανάλυση των προτιμήσεων των χρηστών θα μπορούσαμε να αναδιατάξουμε το μενού της πύλης. Η αναδιάταξη αυτή για παράδειγμα θα μπορούσε να ταξινομήσει τις επιλογές ενός μενού ή υπομενού όχι με αλφαβητική σειρά αλλά με τις προτιμότερες επιλογές να βρίσκονται ψηλότερα στην κάθε λίστα.

Σενάριο	Αριθμός ομάδων
Σενάριο 1	3 ομάδες
Σενάριο 2	2 ομάδες
Σενάριο 3	3 ομάδες
Σενάριο 4	3 ομάδες

Παρατηρώντας τους πίνακες με τα αποτελέσματα της συσταδοποίησης στο προηγούμενο κεφάλαιο βλέπουμε ότι σε όλες τις χρονικές περιόδους, αλλά και στο μεγαλύτερο ποσοστό των συστάδων που υπολογίστηκαν σε όλες τις περιπτώσεις χαρακτηριστικών, ότι οι συστάδες αυτές είναι πολύ πολωμένες. Στο δεύτερο σενάριο τα αποτελέσματα διαφοροποιούνται σε σχέση με τα άλλα τρία σενάρια, καθώς η εξεταζόμενη περίοδος χωρίζεται σε περίοδο με υψηλή και χαμηλή επισκεψιμότητα. Στα άλλα τρία σενάρια παρατηρείται και μία τρίτη περίοδος με μέτρια επισκεψιμότητα.

Για την περεταίρω βελτίωση της ίδιας της ιστοσελίδας απαιτούνται δεδομένα που αφορούν τα χαρακτηριστικά του επισκέπτη της ιστοσελίδας, όπως είναι οι σελίδες που προτιμάει. Με τον τρόπο αυτό υπάρχει δυνατότητα πέρα από την βελτίωση της διαφημιστικής καμπάνιας, να βελτιωθεί και να γίνει πιο λειτουργική η ίδια η ιστοσελίδα.

Τα αποτελέσματα έδειξαν ότι οι μήνες Σεπτέμβριος και Αύγουστος είναι «νεκροί μήνες» και για το λόγο αυτό η επιχείρηση μπορεί να μειώσει τις διαφημιστικές της δαπάνες. Τα ποσά που θα μπορέσει να εξοικονομήσει από την παραπάνω κίνηση μπορεί να τα χρησιμοποιήσει στους μήνες που παρατηρείται μεγαλύτερο ενδιαφέρον.

Βιβλιογραφία

Agresti, A.,Categorical Data Analysis. John Wiley & Sons, Inc., New York,1990.

Akaike, H., A new look at statistical model identification. IEEE Transactions on Automatic Control 19, 1974.

Azzalini, A., Statistical Inference: An Introduction Based on the Likelihood Principle. Springer-Verlag, 1992.

Ansari, S., R. Kohavi, L. Mason, Z. Zheng, Integrating e-commerce and data mining: Architecture and challenges, in: N. Cercone, T.Y. Lin, X. Wu (Eds.), Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM 2001), IEEE Computer Society, 2001.

Banerjee, A., J. Ghosh, Clickstream clustering using weighted longest common subsequences, in: Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining, 2001.

Berendt, B., B. Mobasher, M. Nakagawa, M. Spiliopoulou, The impact of site structure and user environment on session reconstruction in web usage analysis, in: Proceedings of the 4th WebKDD 2002 Workshop, at the ACM-SIGKDD Conference on Knowledge Discovery in Databases, 2002.

Berry, M. and Linoff, G. ,Data Mining Techniques for Marketing, Sales, and Customer Support. John Wiley & Sons, Inc., New York, 1997.

Berry, M. and Linoff, G., Mastering Data Mining. John Wiley & Sons, Inc., New York, 2000.

Berry, M. A. and Linoff. G., Mining the Web: Transforming Customer Data. John Wiley & Sons, Inc., New York, 2002.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A., Discovering Data Mining: From Concept to Implementation. Prentice Hall, Englewood Cliffs NJ, 1997.

Castelo, R. and Giudici, P., Improving Markov Chain model search for data mining. Machine Learning. 2003.

Cooley, R.,B. Mobasher, J. Srivastava, Data preparation for mining world wide web

browsing patterns, Knowledge and Information Systems 1 (1), 1999.

Cooley, R., Web usage mining: discovery and application of interesting patterns from web data, Ph.D. thesis, University of Minnesota, 2000.

Ester, M., H-P. Kriegel, J. Sander and X. Xu. "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases", Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996, pp 226-231.

Fenstermacher, K.D., M. Ginsburg, Mining client-side activity for personalization, in: Fourth IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems (WECWIS_02), 2002, pp. 205–212.

Ferenc B. A trie-based apriori implementation for mining frequent item sequences. In OSDM '05: Proceedings of the 1st international workshop on open source data mining, pages 56–65, New York, NY, USA, 2005. ACM.

Greene, W. H., Econometric Analysis. Prentice Hall, New York, 1999.

Guha, S., Rastogi R. and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases", Proceedings of the ACM SIGMOD Conference, 1998.

Guha, S. Rastogi and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes", Proceedings of the IEEE Conference on Data Engineering, 1999.

Han, J. and Kamber, M., Data Mining: Concepts and Techniques. Morgan Kaufmann, New York, 2001.

Hand, D., Construction and Assessment of Classification Rules. John Wiley & Sons, Ltd, Chichester, 1997.

Heer, J., E.H. Chi, Mining the structure of user activity using cluster stability, in: Proceedings of the Workshop on Web Analytics, Second SIAM Conference on Data Mining, ACM Press, 2002.

T. Kanungo, D. M. Mount, N. S. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans. Patt. Anal. Mach. Intell., 24, 2002.

Karypis, G, E.-H. Han and V. Kumar, “Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modelling”, IEEE Computer, Vol, 32, Issue 8, 1999, pp 68-75.

Kloesgen, W. and Zytkow, J., Handbook of Data Mining and Knowledge Discovery. Oxford University Press, Oxford, 2002.

Ming-Syan Chen, Jong Soo Park, and Philip S. Yu. Efficient data mining for path traversal patterns. IEEE Trans. on Knowl. and Data Eng., 10(2):209–221, 1998.

Mobasher, B., R. Cooley, J. Srivastava, Automatic personalization based on web usage mining, Communications of the ACM 43 (8), 2000.

Nan Niu, E.S., M. El-Ramly, Understanding web usage for dynamic web-site adaptation: A case study, in: Proceedings of the Fourth International Workshop on Web Site Evolution (WSE_02), IEEE, 2002

Nanopoulos, A., D. Katsaros, Y. Manolopoulos, Exploiting web log mining for web cache enhancement, in: R.Kohavi, B. Masand, M. Spiliopoulou, J. Srivastava (Eds.), WEBKDD 2001—Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2011. Revised papers, vol. 2356 of Lecture Notes in Computer Science, Springer, 2002

Kamal Nigam, Andrew McCallum, Tom M. Mitchell: Text Classification from Labeled and Unlabeled Documents using EM, 2000

Oikonomakou, N. and M. Vazirgiannis: “A Review of Web Document Clustering Approaches”, The Data Mining and Knowledge Discovery Handbook, (Edited by O. Maimon and L. Rokach), 2005, pp 921-943.

Pavel Berkhin, Survey Of Clustering Data Mining Techniques, Technical Report, Accrue Software, 2002.

Ramesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, Proceedings of the 1993 ACM SIGMOD International Conference on Manage- ment of Data, pages 207–216, Washington, D.C., FebruaryJune–FebruaryAugust˜ 1993.

Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent item sets. Journal of Parallel and Distributed

Computing, 61(3), 2001.

Sander, K., Scheich, H., 1998. Auditory perception of laughing and crying activates human amygdala regardless of attentional state. Brain Res. Cogn. Brain Res. 12, 181 – 198.

Shahabi, C., F.Banaei Kashani, A framework for efficient and anonymous web usage mining based on client-side tracking, in: R. Kohavi, B. Masand, M. Spiliopoulou, J. Srivastava (Eds.), WEBKDD 2001—Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001. Revised papers, vol. 2356 of Lecture Notes in Computer Science, Springer, 2002

Zanasi, A. (ed.) Text Mining and Its Applications. WIT Press, Southampton, 2003.

Zhang, T., Ramakrishnan R. and Livny, M. "BIRCH: an Efficient Data Clustering Method for Very Large Databases", Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Canada, 1996, pp 103-114.