ΑΛΕΞΑΝΔΡΕΙΟ ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ ΕΜΠΟΡΙΑΣ ΚΑΙ ΔΙΑΦΗΜΙΣΗΣ

Πτυχιακή Εργασία:

Τεχνικές εξόρυξης δεδομένων σε εφαρμογές CRM

Σπουδαστές:
Δερμεντζής Ηλίας
Οικονόμου Μάριος

Επιβλέπων Καθηγητής:
Ασημακόπουλους Κωνσταντίνος

Οκτώβριος 2012

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Κεφάλαιο 3

Εξόρυξη Δεδομένων στα συστήματα CRM

Κεφάλαιο 4

Συμπεράσματα-Προτάσεις

Πρόλογος

Το CRM (Customer Relationship Management) ή στα ελληνικά, Διαχείριση Πελατειακών Σχέσεων, αποτελεί "μία επιχειρησιακή στρατηγική που σχεδιάστηκε με κύριο σκοπό της να βοηθήσει τις επιχειρήσεις να γνωρίσουν τους υπάρχοντες ή πιθανούς πελάτες τους και να δημιουργήσουν ισχυρές πελατειακές σχέσεις με την πάροδο του χρόνου". (Δ. Κοσμάτος, 2011)

Για να γνωρίσει όμως η επιχείρηση τους πελάτες της χρειάζεται πληροφορίες για αυτούς. Για την εξεύρεση και ανάλυση των στοιχείων που διαθέτει ή αποκτά η επιχείρηση, χρησιμοποιούνται οι τεχνικές εξόρυξης δεδομένων (data-mining techniques). Η εξόρυξη δεδομένων είναι ένα "εργαλείο" αυτόματης εξαγωγής σημαντικών «κρυμμένων» πληροφοριών από μια βάση δεδομένων που επιτρέπει στην επιχείρηση να φιλτράρει και να ομαδοποιήσει τους πελάτες της, να προβλέψει κινδύνους ή ευκαιρίες και σαν βασική υπηρεσία να παρέχει πληροφορίες και χαρακτηριστικά γνωρίσματα της συμπεριφοράς των πελατών της.

- **Ο γενικός στόχος** της συγκεκριμένης πτυχιακής εργασίας είναι να παρουσιαστούν οι τεχνικές εξόρυξης δεδομένων που χρησιμοποιούνται στις εφαρμογές CRM και να γίνει κατανοητή η χρήση τους.
- **Οι ειδικοί στόχοι** της εργασίας είναι ο εντοπισμός και η ανάλυση των τεχνικών της εξόρυξης δεδομένων, η επεξήγηση των αλγορίθμων και το πώς χρησιμοποιούνται στην εξόρυξη δεδομένων, οι εφαρμογές τους σε επιχειρήσεις, ο εντοπισμός των τεχνικών που χρησιμοποιούνται στο CRM και οι προτάσεις για βελτίωση.

Αρχικά γίνεται μια βιβλιογραφική επισκόπηση της θεωρίας της εξόρυξης δεδομένων όπως οι εργασίες που εκτελούνται με την εξόρυξη δεδομένων, ο Ενάρετος κύκλος και η Μεθοδολογία της εξόρυξης δεδομένων.

Στη συνέχεια αναλύονται οι τεχνικές εξόρυξης δεδομένων όπως τα δέντρα απόφασης (decision trees), τα τεχνητά νευρωνικά δίκτυα (artificial neural networks), η προσέγγιση του κοντινότερου γείτονα (Nearest Neighbor Approaches), η ανάλυση συνδέσεων (link analysis), η συσταδοποίηση (Clustering) και οι γενετικοί αλγόριθμοι (genetic algorithms) και τέλος πως γίνεται η προετοιμασία των δεδομένων για εξόρυξη.

Έπειτα εξηγούμε πως χρησιμοποιούνται οι τεχνικές εξόρυξης δεδομένων στις εφαρμογές CRM και τι παρέχουν σε μία επιχείρηση.

Στο τέλος αναλύονται τα συμπεράσματα που βγάλαμε από την εργασία και οι όποιες προτάσεις έχουμε να κάνουμε.

Κεφάλαιο 1
Εισαγωγή

Τι είναι η εξόρυξη δεδομένων

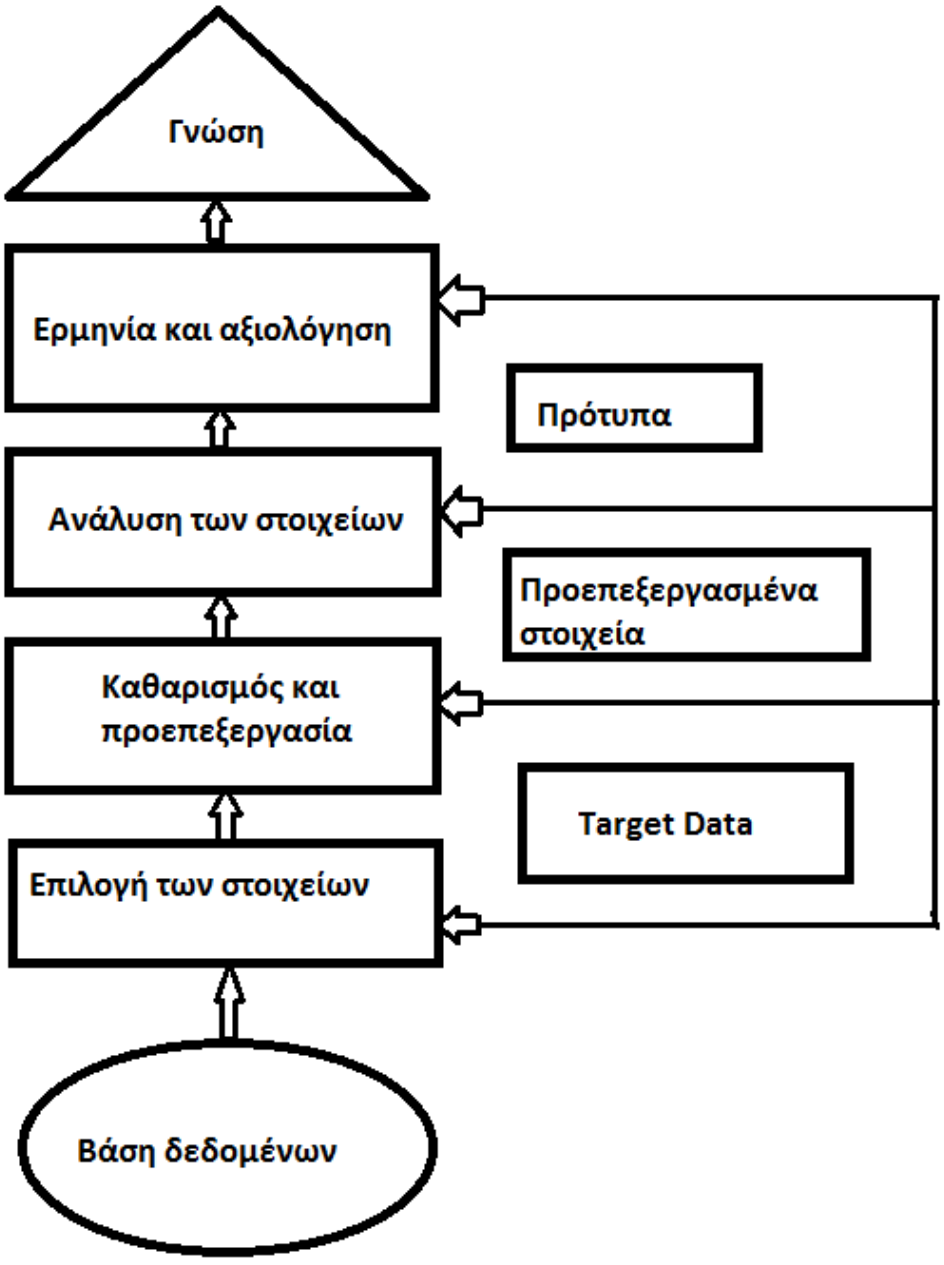
Η εξόρυξη δεδομένων είναι η εξερεύνηση και η ανάλυση μεγάλων ποσοτήτων στοιχείων προκειμένου να ανακαλυφθούν σημαντικά σχέδια και κανόνες. Ένας σκοπός ,και στόχος της εξόρυξης δεδομένων είναι η βελτίωση του μάρκετινγκ, των πωλήσεων, και των διαδικασιών υποστήριξης πελατών, μέσω της καλύτερης κατανόησης των πελατών. Επίσης αυτές οι τεχνικές και τα εργαλεία εξόρυξης δεδομένων χρησιμοποιούνται εξίσου σε τομείς όπως την επιβολή νόμου, την αστρονομία, την ιατρική, και τον έλεγχο βιομηχανικών διεργασιών. Αρχικά, οι αλγόριθμοι εξόρυξης δεδομένων δεν εφευρέθηκαν με σκοπό τις εμπορικές εφαρμογές.

Η εξόρυξη δεδομένων έχει δυο μορφές, την κατευθυνόμενη και την μη κατευθυνόμενη. Η κατευθυνόμενη εξόρυξη δεδομένων προσπαθεί να εξηγήσει ή να ταξινομήσει κάποιο ιδιαίτερο τομέα στόχων όπως για παράδειγμα το εισόδημα. Αντίθετα η μη κατευθυνόμενη εξόρυξη δεδομένων προσπαθεί να βρει τα σχέδια ή τις ομοιότητες μεταξύ των ομάδων-αρχείων χωρίς τη χρήση ενός ιδιαίτερου τομέα στόχων ή κάποιας συλλογής προκαθορισμένων κατηγοριών.

Η εξόρυξη δεδομένων ασχολείται κατά ένα μεγάλο μέρος με την οικοδόμηση προτύπων. Συγκεκριμένα ένα πρότυπο είναι απλά ένας αλγόριθμος ή ένα σύνολο κανόνων που συνδέουν μια συλλογή στοιχείων με έναν ιδιαίτερο στόχο.

Κατ’ επέκταση η οπισθοδρόμηση, τα νευρωνικά δίκτυα, τα δέντρα απόφασης, και οι περισσότερες από τις άλλες τεχνικές εξόρυξης δεδομένων είναι τεχνικές δημιουργίας προτύπων. Κάτω από τις σωστές περιστάσεις, ένα πρότυπο μπορεί να οδηγήσει στην επίγνωση παρέχοντας εξηγήσεις για αποτελέσματα που έχουν σχέση με το πρόβλημα, όπως την τοποθέτηση μιας παραγγελίας ή την αποτυχία να πληρωθεί ένας λογαριασμός, είναι σχετικές και μπορούν να προβλεφτούν από τα διαθέσιμα στοιχεία. Επίσης τα πρότυπα χρησιμοποιούνται για να παράγουν αποτελέσματα. Συγκεκριμένα τα αποτελέσματα είναι ο τρόπος να εκφράζουμε τα ευρήματα ενός προτύπου με έναν ενιαίο αριθμό. Με αυτόν τον τρόπο τα αποτελέσματα μπορούν να χρησιμοποιηθούν για παράδειγμα στην ταξινόμηση ενός κατάλογου πελατών από τους ποιο πιστούς στους λιγότερο πιστούς ή από τους περισσότερο πιθανό να ανταποκριθούν στους λιγότερο.

Τέλος η διαδικασία εξόρυξης δεδομένων αναφέρεται μερικές φορές ως ανακάλυψη γνώσης ή KDD (knowledge discovery in databases). Επίσης θα μπορούσαμε να το σκεφτούμε και ως δημιουργία γνώσης (Berry and Linoff, 2004).



Σχήμα 1. Μια επισκόπηση της διαδικασίας εξόρυξης δεδομένων
(U. Fayyad, G. Piatetsky, P. Smyth, 1996)

Κεφάλαιο 2

Βιβλιογραφική επισκόπηση της θεωρίας της εξόρυξης δεδομένων

2.1 Ποιες εργασίες μπορούν να εκτελεσθούν με την εξόρυξη δεδομένων;

Πολλά προβλήματα διανοητικού, οικονομικού, και επιχειρησιακού συμφέροντος μπορούν να διατυπωθούν από την άποψη των ακόλουθων έξι εργασιών:

- ταξινόμηση
- εκτίμηση
- πρόβλεψη
- ομαδοποίηση συγγένειας
- συγκέντρωση
- περιγραφή και σκιαγράφηση

Τα πρώτα τρία είναι παραδείγματα της κατευθυνόμενης εξόρυξης δεδομένων, όπου ο στόχος είναι να βρεθεί η αξία μιας ιδιαίτερης μεταβλητής στόχων. Αντίθετα η ομαδοποίηση συγγένειας όπως επίσης και η συγκέντρωση είναι παραδείγματα μη κατευθυνόμενης εξόρυξης δεδομένων και στόχος είναι να αποκαλυφθεί η δομή στα στοιχεία χωρίς να δοθεί βάση σε μια μεταβλητή στόχο. Τέλος η σκιαγράφηση είναι ένας περιγραφικός στόχος που μπορεί να είναι είτε κατευθυνόμενος είτε μη κατευθυνόμενος.

2.1.1 Ταξινόμηση

Η ταξινόμηση, είναι ένας από τους πιο κοινούς στόχους της εξόρυξης δεδομένων. Συγκεκριμένα η ταξινόμηση αποτελείται από την εξέταση των χαρακτηριστικών γνωρισμάτων ενός πρόσφατα παρουσιασμένου αντικειμένου και την κατάταξη του σε ένα προκαθορισμένο σύνολο κατηγοριών. Τα αντικείμενα που ταξινομούνται αντιπροσωπεύονται γενικά από τα αρχεία σε έναν πίνακα βάσεων δεδομένων ή ένα αρχείο, και η πράξη της ταξινόμησης αποτελείται από την προσθήκη μιας νέας στήλης με έναν κώδικα κατηγορίας κάποιου είδους. Επίσης η ταξινόμηση χαρακτηρίζεται από ένα σαφή καθορισμό των κατηγοριών, και ένα σύνολο δεδομένων προς εκπαίδευση που αποτελείται από τα παραδείγματα. Ο στόχος είναι να χτιστεί ένα πρότυπο κάποιου είδους που μπορεί να εφαρμοστεί στα αταξινόμητα στοιχεία προκειμένου να τα ταξινομήσει.

Μερικά παραδείγματα των στόχων ταξινόμησης περιλαμβάνουν:

- Ταξινόμηση των πιστωτικών υποψηφίων σε χαμηλού, μέσου, ή υψηλού κίνδυνου
- Επιλογή του περιεχόμενου που εμφανίζετε σε μια ιστοσελίδα
- Καθορισμός τηλεφωνικών αριθμών που αντιστοιχούν στις ανάλογες μηχανές fax
- Επισήμανση των ψευδών ασφαλιστικών αξιώσεων

2.1.2 Εκτίμηση

Η ταξινόμηση ασχολείται με διακριτά αποτελέσματα στην ουσία εξετάζει τις ιδιαίτερες εκβάσεις. Αντίθετα η εκτίμηση εξετάζει τις συνεχώς εκτιμημένες εκβάσεις. Ως εκ τούτου λαμβάνοντας υπόψη μερικά δεδομένα εισόδου, η εκτίμηση βρίσκει μια αξία για κάποια άγνωστη συνεχή μεταβλητή όπως το εισόδημα, το ύψος, ή την ισορροπία πιστωτικών καρτών.

Στην πράξη, η εκτίμηση χρησιμοποιείται συχνά για να εκτελέσει έναν στόχο ταξινόμησης. Τα παραδείγματα των στόχων εκτίμησης περιλαμβάνουν:

- Υπολογισμός του αριθμού παιδιών σε μια οικογένεια
- Υπολογισμός του συνολικού εισοδήματος μιας οικογένειας
- Υπολογισμός της αξίας διάρκειας ζωής ενός πελάτη

2.1.3 Πρόβλεψη

Η πρόβλεψη είναι παρόμοια με την ταξινόμηση ή την εκτίμηση, εκτός από το ότι τα αρχεία είναι ταξινομημένα σύμφωνα με κάποια προσληφθείσα μελλοντική συμπεριφορά ή κατ' εκτίμηση μελλοντική αξία. Σε έναν προβλεπόμενο στόχο, ο μόνος τρόπος να ελεγχθεί η ακρίβεια της ταξινόμησης είναι να περιμένουμε και να δούμε. Επομένως ο αρχικός λόγος που η πρόβλεψη είναι χωρισμένη από την ταξινόμηση και την εκτίμηση είναι ότι στο προβλεπτικό πρότυπο υπάρχουν πρόσθετα ζητήματα σχετικά με τη χρονική σχέση των μεταβλητών.

Οποιεσδήποτε από τις τεχνικές που χρησιμοποιούνται για την ταξινόμηση και την εκτίμηση μπορούν να προσαρμοστούν και στην πρόβλεψη με τη χρησιμοποίηση των παραδειγμάτων εκπαίδευσης όπου η αξία της μεταβλητής που προβλέπεται είναι ήδη γνωστή, μαζί με τα ιστορικά στοιχεία. Έπειτα το ιστορικό στοιχείο χρησιμοποιείται για να χτίσει ένα μοντέλο που εξηγεί την τρέχουσα παρατηρηθείσα συμπεριφορά. Τέλος όταν αυτό το μοντέλο εφαρμόζεται στις τρέχουσες εισαγωγές, το αποτέλεσμα είναι μια πρόβλεψη της μελλοντικής συμπεριφοράς.

Τα παραδείγματα των στόχων πρόβλεψης περιλαμβάνουν:

- Την πρόβλεψη για το ποιοι πελάτες θα φύγουν μέσα στους επόμενους 6 μήνες
- Την πρόβλεψη για το ποιοι τηλεφωνικοί συνδρομητές θα επιλέξουν μια υπηρεσία χρέωσης όπως η τριπλή κλήση ή το φωνητικό ταχυδρομείο

Οι περισσότερες από τις τεχνικές εξόρυξης δεδομένων είναι κατάλληλες για πρόβλεψη εφ' όσον το στοιχείο κατάρτισης είναι διαθέσιμο στην κατάλληλη μορφή. Κατ' επέκταση η επιλογή της τεχνικής εξαρτάται από τη φύση των δεδομένων εισόδου, τον τύπο αξίας που προβλέπεται, και τη σημασία που αποδίδεται στο κατά πόσο μπορεί να εξηγηθεί η πρόβλεψη.

2.1.4 Κανόνες ομαδοποίησης ή ένωσης συγγένειας

Ο στόχος της ομαδοποίησης είναι να καθορίσει ποια πράγματα πηγαίνουν μαζί. Όπως για παράδειγμα τον καθορισμό των αντικειμένων που πηγαίνουν μαζί σε ένα καλάθι αγορών στο supermarket, που είναι ο κεντρικός στόχος της ανάλυσης καλαθιών αγοράς. Επίσης οι λιανικές αλυσίδες μπορούν να χρησιμοποιήσουν την ομαδοποίηση για να προγραμματίσουν και να ρυθμίσουν τα αντικείμενα στα ράφια καταστημάτων ή σε έναν κατάλογο έτσι ώστε τα στοιχεία που αγοράζονται συχνά μαζί να βρίσκονται και μαζί. Η ομαδοποίηση συγγένειας μπορεί επίσης να χρησιμοποιηθεί για να προσδιορίσει τις ευκαιρίες πώλησης όπως επίσης για να σχεδιάσει ελκυστικές συσκευασίες ή πακέτα προϊόντων η και υπηρεσιών.

Η ομαδοποίηση είναι μια απλή προσέγγιση στην δημιουργία κανόνων από στοιχεία. Εάν δύο στοιχεία, για παράδειγμα γάλα και αλεύρι, εμφανίζονται μαζί αρκετά συχνά, μπορούμε να παραγάγουμε δύο κανόνες ένωσης:

- Οι άνθρωποι που αγοράζουν γάλα αγοράζουν επίσης αλεύρι με την πιθανότητα Π1.
- Οι άνθρωποι που αγοράζουν αλεύρι αγοράζουν επίσης γάλα με την πιθανότητα Π2.

2.1.5 Συγκέντρωση

Η συγκέντρωση είναι η εργασία κατάτμησης ενός ετερογενή πληθυσμού σε διάφορες πιο ομοιογενείς υποομάδες ή συστάδες. Συγκεκριμένα αυτό που διακρίνει τη συγκέντρωση από την ταξινόμηση είναι ότι η συγκέντρωση δεν στηρίζεται στις προκαθορισμένες κατηγορίες. Στην ταξινόμηση, κάθε αρχείο είναι ορισμένο σε μια προκαθορισμένη κατηγορία βάσει ενός μοντέλου. Επίσης στη συγκέντρωση, δεν υπάρχει καμία προκαθορισμένη κατηγορία και κανένα παράδειγμα. Τα αρχεία συγκεντρώνονται βάσει της ομοιότητας.

Η συγκέντρωση χρησιμοποιείται συχνά ως προοίμιο σε κάποια άλλη μορφή εξόρυξης δεδομένων ή διαμόρφωσης. Παραδείγματος χάριν, η συγκέντρωση να είναι το πρώτο βήμα σε μια προσπάθεια κατάτμησης της αγοράς: Αντί της προσπάθειας να βρεθεί ένας κανόνας που θα ισχύει για όλους και όλα, μπορεί αρχικά να διαιρεθεί η βάση πελατών σε συστάδες ή ανθρώπους με παρόμοιες συνήθειες αγοράς, και έπειτα να απαντηθεί ποιο είδος προώθησης λειτουργεί καλύτερα για κάθε συστάδα.

2.1.6 Σκιαγράφηση

Μερικές φορές ο σκοπός της εξόρυξης δεδομένων είναι απλά να περιγραφεί τι συμβαίνει σε μια περίπλοκη βάση δεδομένων με έναν τρόπο που να αυξάνει την κατανόησή μας όσον αφορά τους ανθρώπους, τα προϊόντα, ή τις διαδικασίες που παρήγαγαν τα στοιχεία αρχικά. Μια καλή περιγραφή μιας συμπεριφοράς συχνά θα προτείνει και μια εξήγηση για αυτήν. Στο ελάχιστο, μια καλή περιγραφή μπορεί να προτείνει από πού να αρχίσει η έρευνα για την εξήγηση.

Οι περισσότερες από τις τεχνικές εξόρυξης δεδομένων έχουν υπάρξει, τουλάχιστον ως ακαδημαϊκοί αλγόριθμοι, για χρόνια ή δεκαετίες. Εντούτοις, μόνο την τελευταία δεκαετία η εμπορική εξόρυξη δεδομένων χρησιμοποιείτε με δυναμικό τρόπο. Αυτό οφείλεται στη σύγκλιση διάφορων παραγόντων:

- Τα δεδομένα παράγονται.
- Τα στοιχεία αποθηκεύονται.
- Η δύναμη υπολογισμού είναι προσιτή.
- Το ενδιαφέρον για τη διαχείριση σχέσης πελατών είναι ισχυρό.
- Τα εμπορικά προϊόντα λογισμικού εξόρυξης δεδομένων είναι εύκολα διαθέσιμα.

2.2 Κατευθυνόμενη και μη κατευθυνόμενη εξόρυξη δεδομένων

2.2.1 Κατευθυνόμενη εξόρυξη δεδομένων

Σύμφωνα με τους Berry και Linoff (1999), στην κατευθυνόμενη εξόρυξη δεδομένων (directed data-mining) ο στόχος μας είναι να χρησιμοποιηθούν τα διαθέσιμα δεδομένα ώστε να κατασκευαστεί ένα πρότυπο, το οποίο θα περιγράφει μια συγκεκριμένη μεταβλητή ενδιαφέροντος υπό το πρίσμα των υπόλοιπων διαθέσιμων δεδομένων .

Παραδείγματα κατευθυνόμενης εξόρυξης δεδομένων είναι τα ακόλουθα:

- Ταξινόμηση (classification). Η εξέταση των χαρακτηριστικών ενός αντικειμένου και η τοποθέτησή του σε μια προκαθορισμένη υπάρχουσα κατηγορία.
- Εκτίμηση (estimation). Ο υπολογισμός μιας τιμής για μια άγνωστη συνεχή μεταβλητή, με βάση τα δεδομένα εισόδου.
- Πρόβλεψη (prediction). Η πρόβλεψη, μέσω ενός προτύπου, αγνώστων ή μη διαθέσιμων τιμών.

2.2.2 Η μη-κατευθυνόμενη εξόρυξη δεδομένων.

Στη μη-κατευθυνόμενη ή ελεύθερη εξόρυξη δεδομένων (undirected data-mining), δεν επιλέγεται κάποια μεταβλητή ως στόχος. Ο σκοπός της είναι να δημιουργηθεί κάποια σχέση ανάμεσα σε όλες τις μεταβλητές. Παράδειγμα της μη-κατευθυνόμενης εξόρυξης δεδομένων είναι τα ακόλουθα (Berry and Linoff 1999):

- Ομαδοποίηση ή κανόνες συσχέτισης (affinity grouping and association rules). Ο καθορισμός των αντικειμένων που συμβαδίζουν μαζί.
- Συσταδοποίηση (clustering). Η κατάτμηση μίας ποικιλόμορφης ομάδας αντικειμένων σε συναφείς υποομάδες ή συστάδες (clusters).
- Περιγραφή και απεικόνιση (description and visualization). Η αναλυτική περιγραφή των αντικειμένων μιας βάσης δεδομένων με σκοπό τη βαθύτερη κατανόηση των δεδομένων και την απεικόνισή τους με τρόπο κατανοητό από τον ανθρώπινο παράγοντα.

2.3 Ποιος είναι ο ενάρετος κύκλος;

Σύμφωνα με τους Berry και Linoff (2004) ο ενάρετος κύκλος της εξόρυξης δεδομένων αποτελείτε από τέσσερα στάδια:

1. Προσδιορισμός του επιχειρησιακού προβλήματος.

2.Στοιχεία μεταλλείας/εξόρυξης για να μετασχηματιστούν τα στοιχεία σε αγώγιμες πληροφορίες.

3. Ενέργεια μέσω των πληροφοριών.

4. Μέτρηση των αποτελεσμάτων.

Όπως αυτά τα βήματα προτείνουν, το κλειδί για την επιτυχία είναι η ενσωμάτωση της εξόρυξης δεδομένων εντός των επιχειρησιακών διαδικασιών και η ενθάρρυνση των γραμμών επικοινωνίας μεταξύ των data-miners και των επιχειρησιακών χρηστών των αποτελεσμάτων.

2.3.1 Προσδιορίζοντας την επιχειρησιακή ευκαιρία

Ο ενάρετος κύκλος της εξόρυξης δεδομένων αρχίζει με τον προσδιορισμό των σωστών επιχειρησιακών ευκαιριών. Δυστυχώς, υπάρχουν πάρα πολλοί καλοί στατιστικοί και ικανοί αναλυτές των οποίων η εργασία ουσιαστικά σπαταλιέται επειδή λύνουν προβλήματα που δεν βοηθούν την επιχείρηση. Οι καλοί ανθρακωρύχοι στοιχείων (Data miners) θέλουν να αποφύγουν αυτήν την κατάσταση. Η αποφυγή αυτής της χρονικής σπατάλης ξεκινά με την θέληση να ενεργείς σύμφωνα με τα αποτελέσματα. Πολλές επιχειρήσεις είναι καλές υποψήφιες για εξόρυξη δεδομένων (Berry and Linoff, 2004):

- Προγραμματισμός για μια εισαγωγή νέων προϊόντων.
- Προγραμματισμός των άμεσων εκστρατειών μάρκετινγκ.
- Κατανόηση της τριβής / αποχώρησης των πελατών.
- Αξιολόγηση των αποτελεσμάτων μιας εκστρατείας μάρκετινγκ.

Αυτά είναι παραδείγματα όπου η εξόρυξη δεδομένων μπορεί να ενισχύσει τις υπάρχουσες επιχειρησιακές προσπάθειες, με την άδεια των επιχειρησιακών διευθυντών έτσι ώστε να είναι σε θέση να παίρνουν περισσότερο ενημερωμένες αποφάσεις.

Οι μετρήσεις των προηγούμενων προσπαθειών και των ειδικών ερωτήσεων για την επιχείρηση προτείνουν επίσης ευκαιρίες εξόρυξης δεδομένων:

- Ποιοι τύποι πελατών αποκρίθηκαν στην τελευταία εκστρατεία;
- Πού ζουν οι καλύτεροι πελάτες;

- Ποια τα αιτία της τριβής των πελατών;
- Οι κερδοφόροι πελάτες χρησιμοποιούν την υποστήριξη πελατών;
- Ποια προϊόντα θα έπρεπε να προαχθούν;

2.3.2 Στοιχεία εξόρυξης (Mining data)

Η εξόρυξη δεδομένων, μετασχηματίζει τα στοιχεία σε αγώγιμα αποτελέσματα. Παρόλα αυτά υπάρχουν πολυάριθμες παγίδες που παρεμποδίζουν τη δυνατότητα να χρησιμοποιηθούν τα αποτελέσματα της εξόρυξης δεδομένων:

- Κακές μορφές δεδομένων
- Σύγχυση των τομέων δεδομένων, όπως μια ημερομηνία παράδοσης που σημαίνει τη "προγραμματισμένη ημερομηνία παράδοσης" σε ένα σύστημα και τη "πραγματική ημερομηνία παράδοσης" σε ένα άλλο σύστημα
- Έλλειψη λειτουργίας,
- Νομικές επιπτώσεις, όπως να πρέπει να παρασχεθεί ένας νομικός λόγος κατά την απόρριψη ενός δανείου (και μια απάντηση όπως "το νευρωνικό δίκτυό υπέδειξε έτσι" δεν είναι αποδεκτή)
- Οργανωτικοί παράγοντες, δεδομένου ότι μερικές λειτουργικές ομάδες είναι απρόθυμες να αλλάξουν τις διαδικασίες τους, ιδιαίτερα χωρίς κίνητρα
- Έλλειψη επικαιρότητας, δεδομένου ότι τα αποτελέσματα που έρχονται πάρα πολύ αργά δεν μπορούν πλέον να είναι αγώγιμα

2.3.3 Λαμβάνοντας μέτρα

Η λήψη μέτρων είναι ο σκοπός του ενάρετου κύκλου της εξόρυξης δεδομένων. Όπως αναφέρεται ήδη, η δράση μπορεί να λάβει πολλές μορφές. Συγκεκριμένα η εξόρυξη δεδομένων καθιστά τις επιχειρησιακές αποφάσεις ενημερωμένες. Επίσης κατά τη διάρκεια του χρόνου, αναμένουμε ότι οι καλύτερα-ενημερωμένες αποφάσεις οδηγούν σε καλύτερα αποτελέσματα.

Οι ενέργειες συνήθως πρέπει να συμφωνούν με αυτό που η επιχείρηση κάνει:

- Στέλνοντας τα μηνύματα στους πελάτες και τις προοπτικές μέσω του άμεσου ταχυδρομείου, ηλεκτρονικό ταχυδρομείο κλπ με την εξόρυξη δεδομένων, διαφορετικά μηνύματα μπορούν να πάνε σε διαφορετικούς ανθρώπους
- Να δώσει προτεραιότητα στην εξυπηρέτηση πελατών
- Ρυθμίζοντας επίπεδα καταλόγων

Τα αποτελέσματα της εξόρυξης δεδομένων πρέπει να τροφοδοτούν τις επιχειρησιακές διαδικασίες που αγγίζουν τους πελάτες και έχουν επιπτώσεις στη σχέση με τους πελάτες. (Daniel T. Larose, 2001)

2.3.4 Μέτρηση των αποτελεσμάτων

Η σημασία της μέτρησης αποτελεσμάτων έχει ήδη τονιστεί. Παρά τη μεγάλη σημασία του, είναι το στάδιο του ενάρετου κύκλου που πιθανότατα αγνοείται. Ακόμα κι αν η αξία της μέτρησης και της συνεχούς βελτίωσης αναγνωρίζεται ευρέως, δίνεται συνήθως λιγότερη προσοχή από αυτή που του αξίζει. Για παράδειγμα πολλές επιχειρήσεις είναι έτοιμες να εφαρμόσουν προγράμματα και μέτρα χωρίς να ελέγξουν εάν το πλάνο τους αντιστοιχεί στην πραγματικότητα.

Ως εκ τούτου ο κατάλληλος χρόνος για να αρχίσουν οι μετρήσεις είναι στην αρχή, κατά τον προσδιορισμό του επιχειρησιακού προβλήματος.

Μερικά παραδείγματα ερωτήσεων τα οποία μπορούν να βοηθήσουν στην μέτρηση και να δώσουν μελλοντικά αποτελέσματα είναι:

- Η εκστρατεία κατάφερε να προσεγγίσει και να φέρει κερδοφόρους πελάτες;
- Αυτοί οι πελάτες διατηρήθηκαν όπως αναμένονταν;
- Ποια είναι τα χαρακτηριστικά των πιστών πελατών που επιτυγχάνονται από αυτήν την εκστρατεία; Τα δημογραφικά σχεδιαγράμματα των γνωστών πελατών μπορούν να εφαρμοστούν στους μελλοντικούς ενδεχόμενους πελάτες.
- Οι πελάτες αγοράζουν επιπρόσθετα προϊόντα; Μπορούν τα διαφορετικά συστήματα σε μια οργάνωση να ανιχνεύσουν εάν ένας πελάτης αγοράζει πολλαπλάσια προϊόντα;
- Οι πελάτες που προσεγγιστήκαν από την εκστρατεία αποκρίθηκαν στα εναλλασσόμενα κανάλια;

Όλες αυτές οι μετρήσεις παρέχουν τις πληροφορίες που χρειάζονται για να λαμβάνονται ενημερωμένες αποφάσεις στο μέλλον. Η εξόρυξη δεδομένων αφορά την σύνδεση του παρελθόντος μέσω της εκμάθησης με τις μελλοντικές ενέργειες.

2.4. Μεθοδολογία εξόρυξης δεδομένων και πρακτικές

2.4.1 Μεθοδολογία

Η εξόρυξη δεδομένων είναι ένας τρόπος απόκτησης γνώσης από το παρελθόν ώστε να ληφθούν καλύτερες αποφάσεις στο μέλλον. Οι πρακτικές που περιγράφονται σε αυτό το κεφάλαιο έχουν ως σκοπό να αποφευχθούν δύο ανεπιθύμητες εκβάσεις της διαδικασίας εκμάθησης:

- στοιχειά εκμάθησης που δεν είναι αληθινά
- στοιχειά εκμάθησης που είναι αληθινά, αλλά μη χρήσιμα

Επομένως οι ανθρακωρύχοι στοιχείων πρέπει να ξέρουν πώς να αποφύγουν τους κοινούς κινδύνους. (Daniel T. Larose, 2001)

2.4.1.α Στοιχεία εκμάθησης που δεν είναι αληθινά

Η εκμάθηση των στοιχείων που δεν είναι αληθινά είναι πιο επικίνδυνη από τα στοιχειά εκμάθησης που είναι άχρηστα και αυτό γιατί οι σημαντικές επιχειρησιακές αποφάσεις μπορεί να ληφθούν βασισμένες σε ανακριβείς πληροφορίες. Τα αποτελέσματα εξόρυξης δεδομένων φαίνονται συχνά αξιόπιστα επειδή είναι βασισμένα σε πραγματικά στοιχεία κατά τρόπο φαινομενικά επιστημονικό. Επομένως αυτή η αξιοπιστία μπορεί να είναι παραπλανητική. Επίσης τα ίδια τα στοιχεία μπορούν να είναι ανακριβή ή μη σχετικά με την ερώτηση. Επομένως τα σχέδια που ανακαλύπτονται μπορεί να απεικονίζουν προηγούμενες επιχειρησιακές αποφάσεις ή και τίποτα. Οι μετασχηματισμοί στοιχείων όπως η περιληπτική παρουσίαση της πληροφορίας μπορεί να έχουν καταστρέψει ή να έχουν αποκρύψει σημαντικές πληροφορίες. Μερικά από τα πιο κοινά προβλήματα που μπορούν να οδηγήσουν σε λάθος συμπεράσματα είναι:

- Τα σχέδια μπορεί να μην αντιπροσωπεύσουν οποιοδήποτε υποκειμενικό κανόνα.
- Το καθορισμένο πρότυπο μπορεί να μην απεικονίζει τον σχετικό πληθυσμό.
- Τα στοιχεία μπορεί να είναι σε λανθασμένο επίπεδο λεπτομέρειας.

2.4.1.β Στοιχεία εκμάθησης που είναι αληθινά, αλλά μη χρήσιμα

Αν και όχι τόσο επικίνδυνο όσο τα στοιχειά εκμάθησης που δεν είναι αληθινά, τα στοιχεία εκμάθησης που δεν είναι χρήσιμα είναι ένα ποιό κοινό φαινόμενο.
Μερικά παραδείγματα είναι:

- Μαθαίνοντας στοιχεία που είναι είδη γνωστά

- Μαθαίνοντας στοιχεία που δεν μπορούν να χρησιμοποιηθούν

Η μεθοδολογία εξόρυξης δεδομένων έχει ως σκοπό να ξεκαθαρίσει και να κάνει σαφές ποιά στοιχειά είναι χρήσιμα και πια αληθινά. Ως εκ τούτου η μεθοδολογία έχει ως σκοπό να εξασφαλίσει ότι οι προσπάθειες εξόρυξης δεδομένων θα οδηγήσουν σε ένα σταθερό πρότυπο το οποίο θα εξετάζει επιτυχώς την λύση του επιχειρησιακού προβλήματος.

2.4.2 Δοκιμή υπόθεσης

Η δοκιμή υπόθεσης είναι η απλούστερη προσέγγιση στην ενσωμάτωση των στοιχείων στις διαδικασίες λήψης αποφάσεων μιας επιχείρησης. Ο σκοπός της δοκιμής υπόθεσης είναι να τεκμηριωθούν ή να ανασκευαστούν οι ιδέες, και είναι ένα μέρος του συνόλου των προσπαθειών εξόρυξης δεδομένων. Επίσης η δοκιμή υπόθεσης είναι πολυτιμότερη όταν αποκαλύπτει ότι οι υποθέσεις που έχουν καθοδηγήσει τις ενέργειες μιας επιχείρησης στην αγορά είναι ανακριβείς.

Τα βήματα της διαδικασίας δοκιμής υπόθεσης είναι δύο:

- Παραγωγή υποθέσεων
- Εξέταση υποθέσεων

2.4.3. Πρότυπα, σκιαγράφηση, και πρόβλεψη

Η δοκιμή υπόθεσης είναι βεβαίως χρήσιμη, αλλά έρχεται μία στιγμή που δεν είναι πλέον ικανοποιητική. Οι τεχνικές εξόρυξης δεδομένων σχεδιάζονται για την εκμάθηση των νέων πραγμάτων με τη δημιουργία προτύπων βασισμένων στα στοιχεία.

Υπό τη γενικότερη έννοια, ένα πρότυπο είναι μια εξήγηση ή μια περιγραφή για το πώς κάτι λειτουργεί και που απεικονίζει την πραγματικότητα αρκετά καλά έτσι ώστε να μπορεί να διεξαγάγει συμπεράσματα για τον πραγματικό κόσμο.

Οι τεχνικές εξόρυξης δεδομένων μπορούν να χρησιμοποιηθούν για να κάνουν τρία είδη προτύπων για τρία είδη εργασιών: περιγραφική σκιαγράφηση, κατευθυνόμενη σκιαγράφηση, και πρόβλεψη. Οι διακρίσεις δεν είναι πάντα σαφείς. (Berry and Linoff, 2004).

Σκιαγράφηση

Η σκιαγράφηση είναι μια γνωστή προσέγγιση για πολλά προβλήματα και δεν περιλαμβάνει κάποια περίπλοκη ανάλυση στοιχείων. Οι έρευνες, παραδείγματος χάριν, είναι μια κοινή μέθοδος σκιαγράφησης πελατών. Συγκεκριμένα οι έρευνες αποκαλύπτουν ποιοι πελάτες και ποιες προοπτικές μοιάζουν μεταξύ τους, ή τουλάχιστον με ποιον τρόπο απαντώνται οι ερωτήσεις της έρευνας.

Τα σχεδιαγράμματα είναι συχνά βασισμένα στις δημογραφικές μεταβλητές, όπως την γεωγραφική θέση, το φύλο, και την ηλικία. Δεδομένου ότι η διαφήμιση πωλείται σύμφωνα με αυτές τις μεταβλητές, τα δημογραφικά προφίλ μπορούν να μετατραπούν άμεσα σε στρατηγικές των μέσων ενημέρωσης.

Αν και ισχυρή, η σκιαγράφηση έχει σοβαρούς περιορισμούς. Ένας περιορισμός είναι η ανικανότητα να διακριθεί το αίτιο και το αποτέλεσμα. Επίσης εφ' όσον η σκιαγράφηση βασίζεται στις γνωστές δημογραφικές μεταβλητές, δεν είναι εύκολα αντιληπτή. Για παράδειγμα εάν οι άνδρες αγοράζουν μπύρα περισσότερο από ότι οι γυναίκες, δεν χρειάζεται να αναρωτηθούμε το αν η κατανάλωση μπύρας θα μπορούσε να είναι αιτία ανδρισμού. Ως εκ τούτου φαίνεται ασφαλές να υποθέσουμε ότι η σύνδεση είναι από τους άνδρες για την μπύρα και όχι το αντίστροφο. (Rygielski C., 2002)

Μερικά άλλα παραδείγματα από τα πραγματικά προγράμματα εξόρυξης δεδομένων είναι:

- Οι άνθρωποι που έχουν αγοράσει πιστοποιητικά καταθέσεων έχουν ελάχιστα χρήματα αποταμιευμένα.
- Οι πελάτες που χρησιμοποιούν το φωνητικό ταχυδρομείο κάνουν σύντομες κλήσεις στον αριθμό τους.

Πρόβλεψη

Η σκιαγράφηση χρησιμοποιεί τα στοιχεία από το παρελθόν για να περιγράψει τι συνέβη στο παρελθόν. Η πρόβλεψη πηγαίνει ένα βήμα παραπέρα. Συγκεκριμένα η πρόβλεψη χρησιμοποιεί τα στοιχεία από το παρελθόν για να προβλέψει αυτό που είναι πιθανό να συμβεί στο μέλλον. Αυτή η χρήση των στοιχείων είναι ισχυρότερη.

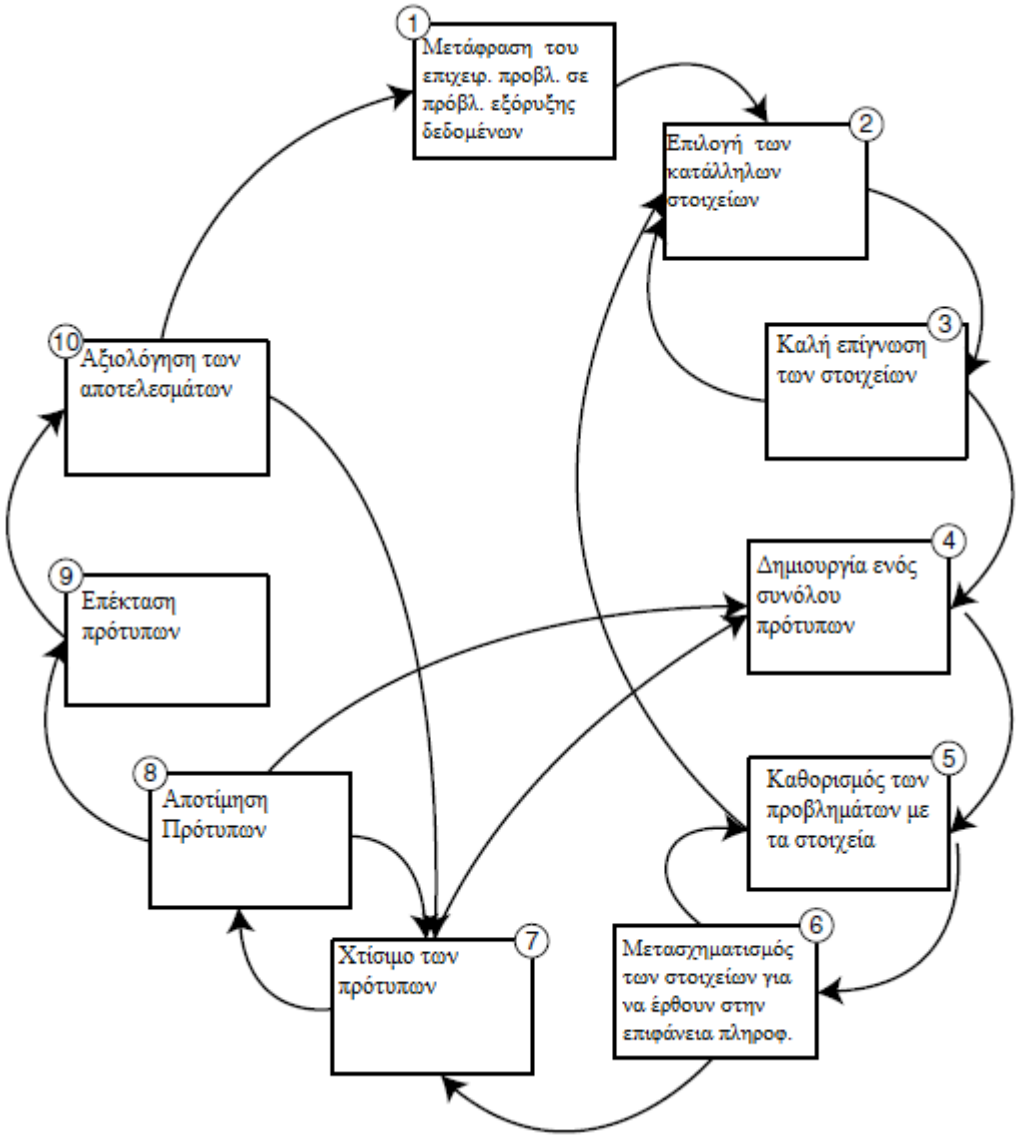
Η οικοδόμηση ενός προβλεπτικού μοντέλου απαιτεί τον έγκαιρο διαχωρισμό ανάμεσα στα πρότυπα εισαγωγών ή των προάγγελων (predictors) και στην παραγωγή του προτύπου. Επομένως εάν αυτός ο διαχωρισμός δεν διατηρηθεί, το πρότυπο δεν θα λειτουργήσει. Αυτό είναι ένα παράδειγμα του λόγου για τον οποίο είναι σημαντικό να ακολουθηθεί μια υγιής μεθοδολογία εξόρυξης δεδομένων.

2.4.4 Η μεθοδολογία

Η μεθοδολογία εξόρυξης δεδομένων έχει 11 βήματα :

- **1.** Μετάφραση του επιχειρησιακού προβλήματος σε πρόβλημα εξόρυξης δεδομένων.
- **2.** Επιλογή των κατάλληλων στοιχείων.
- **3.** Καλή επίγνωση των στοιχείων.
- **4.** Δημιουργία ενός συνόλου πρότυπων.

- **5.** Καθορισμός των προβλημάτων με τα στοιχεία.
- **6.** Μετασχηματισμός των στοιχείων για να έρθουν στην επιφάνεια πληροφορίες.
- **7.** Χτίσιμο των πρότυπων.
- **8.** Αποτίμηση πρότυπων.
- **9.** Επέκταση πρότυπων.
- **10.** Αξιολόγηση των αποτελεσμάτων.
- **11.** Επανεκκίνηση.



Σχέδιο 2.1 Η εξόρυξη δεδομένων δεν είναι μία γραμμίκη διαδηκασία
(Berry and Linoff, 2004)

Τα βήματα έχουν μια φυσική τάξη, αλλά δεν είναι απαραίτητο να τελειώσει ένα βήμα πριν κινηθούμε προς το επόμενο. Επίσης αυτά που μαθαίνονται στα πιο πρόσφατα βήματα, μας

αναγκάζουν πολλές φορές να ξαναεπισκεφτούμε τα προηγούμενα. (Tsau Young Lin, Ying Yie, 2008)

2.5 Δέντρα απόφασης

Τα δέντρα απόφασης είναι δημοφιλή και για την ταξινόμηση και για την πρόβλεψη. Σύμφωνα με τους Berry και Linoff (2004) η ελκυστικότητα των δέντρο-βασισμένων μεθόδων οφείλεται κατά ένα μεγάλο μέρος στο γεγονός ότι τα δέντρα απόφασης αντιπροσωπεύουν κανόνες. Αυτοί οι κανόνες μπορούν εύκολα να εκφραστούν στα αγγλικά (ή σε οποιαδήποτε άλλη γλώσσα) έτσι ώστε οι άνθρωποι να μπορούν να τους καταλάβουν όπως επίσης μπορούν να εκφραστούν και σε μια γλώσσα πρόσβασης βάσεων δεδομένων για να ανακτήσουν αρχεία σε μια ιδιαίτερη κατηγορία.

Τα δέντρα απόφασης είναι επίσης χρήσιμα στην διερεύνηση δεδομένων με σκοπό να αποκτήσουν γνώση που αφορά τις σχέσεις ενός μεγάλου αριθμού υποψήφιων μεταβλητών εισαγωγής σε μια μεταβλητή στόχο.

Επειδή τα δέντρα απόφασης συνδυάζουν την εξερεύνηση και τη διαμόρφωση στοιχείων, είναι ένα ισχυρό πρώτο βήμα στη διαδικασία διαμόρφωσης ακόμα και στο χτίσιμο του τελικού προτύπου που χρησιμοποιεί κάποια άλλη τεχνική.

2.5.1 Τι είναι ένα δέντρο απόφασης

Ένα δέντρο απόφασης είναι μια δομή που μπορεί να χρησιμοποιηθεί για να κατανείμει μια μεγάλη συλλογή αρχείων σε διαδοχικά μικρότερα σύνολα αρχείων με την εφαρμογή μιας ακολουθίας απλών κανόνων απόφασης. Συγκεκριμένα με κάθε διαδοχικό τμήμα, τα μέλη των συνόλων που προκύπτουν γίνονται όλο και περισσότερο παρόμοια .

Ένα πρότυπο δέντρου απόφασης αποτελείται από ένα σύνολο κανόνων με σκοπό την διαίρεση ενός μεγάλου ετερογενή πληθυσμού σε μικρότερες, περισσότερο ομοιογενείς ομάδες σε σχέση με μία συγκεκριμένη μεταβλητή στόχο. Συνήθως η μεταβλητή στόχος είναι κατηγορική και το πρότυπο δέντρων απόφασης χρησιμοποιείται είτε για να υπολογίσει την πιθανότητα ενός δεδομένου αρχείου να ανήκει σε κάθε μια από τις κατηγορίες, είτε για να ταξινομήσει το αρχείο με την κατάταξη της στην πλέον πιθανή κατηγορία. Τα δέντρα απόφασης μπορούν επίσης να χρησιμοποιηθούν για να υπολογίσουν την αξία μιας συνεχούς μεταβλητής, αν και υπάρχουν άλλες καλύτερες τεχνικές για αυτήν την εργασία. (Berry and Linoff, 2004)

2.5.2 Ταξινόμηση στα δέντρα απόφασης

Ένα δέντρο απόφασης αντιπροσωπεύει μια σειρά ερωτήσεων. Η απάντηση στην πρώτη ερώτηση καθορίζει την ακόλουθη ερώτηση. Συγκεκριμένα, οι αρχικές ερωτήσεις δημιουργούν τις ευρείες κατηγορίες με πολλά μέλη, έπειτα οι συνεχόμενες ερωτήσεις διαιρούν τις ευρείες κατηγορίες σε μικρότερα και μικρότερα σύνολα. Επίσης εάν οι ερωτήσεις επιλέγονται σωστά, μια εκπληκτικά σύντομη σειρά ερωτήσεων είναι αρκετή για να ταξινομήσει ένα εισερχόμενο αρχείο.

Ένα αρχείο εισέρχεται στο δέντρο, στον κόμβο ρίζας. Έπειτα ο κόμβος ρίζας εφαρμόζει μια δοκιμή για να καθορίσει ποιο *κόμβο* το αρχείο θα αντιμετωπίσει έπειτα. Μπορεί να υπάρχουν διαφορετικοί αλγόριθμοι για την αρχική δοκιμή, αλλά ο στόχος είναι πάντα ο ίδιος. Αυτή η διαδικασία επαναλαμβάνεται έως ότου φθάσει το αρχείο σε έναν *κόμβο φύλλων*. Όλα τα αρχεία που καταλήγουν σε ένα δεδομένο φύλλο του δέντρου είναι ταξινομημένα με τον ίδιο τρόπο. Ως εκ τούτου υπάρχει μια μοναδική πορεία από τη ρίζα σε κάθε φύλλο. Αυτή η πορεία είναι μια έκφραση του *κανόνα* που χρησιμοποιείται για να ταξινομήσει τα αρχεία.

Επίσης τα διαφορετικά φύλλα μπορούν να κάνουν την ίδια ταξινόμηση, αν και κάθε φύλλο κάνει την ταξινόμηση για έναν διαφορετικό λόγο.

2.5.3 Εκτίμηση στα δέντρα απόφασης.

Ας υποθέσουμε ότι η σημαντική επιχειρησιακή ερώτηση δεν είναι *ποιος θα αποκριθεί* αλλά *ποιο θα είναι το μέγεθος της επόμενης παραγγελίας του πελάτη*; Το δέντρο απόφασης μπορεί να χρησιμοποιηθεί για να απαντήσει σε αυτήν την ερώτηση. Αν υποθέσουμε ότι το ποσό παραγγελίας είναι μια από τις διαθέσιμες μεταβλητές στο πρότυπο σύνολο, τότε το μέσο μέγεθος παραγγελίας σε κάθε φύλλο μπορεί να χρησιμοποιηθεί ως το εκτιμώμενο μέγεθος παραγγελίας για οποιοδήποτε αταξινόμητο αρχείο που ικανοποιεί τα κριτήρια για εκείνο το φύλλο. Επίσης είναι δυνατό να χρησιμοποιηθεί μια αριθμητική μεταβλητή για να χτιστεί το δέντρο, ένα τέτοιο δέντρο ονομάζεται *δέντρο οπισθοδρόμησης*. Στα δέντρα οπισθοδρόμησης αντί της αύξησης της 'καθαρότητας' μιας κατηγορικής μεταβλητής, κάθε διάσπαση στο δέντρο επιλέγεται ώστε να μειωθεί η διακύμανση στις τιμές της μεταβλητής στόχου μέσα σε κάθε κόμβο.

Τα δέντρα μπορούν να χρησιμοποιηθούν για να υπολογίσουν συνεχείς τιμές, παρόλα αυτά δεν συνιστώνται . Ένας ''εκτιμητής'' (υπολογιστής) δέντρων απόφασης μπορεί μόνο να παράγει τόσες τιμές όσα είναι τα φύλλα στο δέντρο. Επομένως για να υπολογιστεί μια συνεχής μεταβλητή, είναι προτιμότερο να χρησιμοποιηθεί μια συνεχής λειτουργία. Τα

πρότυπα οπισθοδρόμησης και τα νευρωνικά πρότυπα δικτύων είναι γενικά πιο κατάλληλα για την εκτίμηση.

2.5.4 Εξαγωγή των κανόνων από τα δέντρα

Όταν ένα δέντρο απόφασης χρησιμοποιείται κυρίως για να παράγει αποτελέσματα, είναι εύκολο να ξεχαστεί ότι ένα δέντρο απόφασης στην πραγματικότητα είναι μια συλλογή κανόνων. Εάν ένας από τους σκοπούς της εξόρυξης δεδομένων είναι η κατανόηση του προβλήματος, χρήσιμο είναι να μειωθεί ο μεγάλος αριθμός των κανόνων σε ένα δέντρο απόφασης σε μια μικρότερη, πιο κατανοητή συλλογή.

Όταν ένα δέντρο απόφασης χρησιμοποιείται για την παραγωγή αποτελεσμάτων, η κατοχή ενός μεγάλου αριθμού φύλλων είναι συμφέρουσα επειδή κάθε φύλλο παράγει ένα διαφορετικό αποτέλεσμα. Επίσης όταν ο στόχος είναι να παραχθούν κανόνες, όσο λιγότεροι είναι οι κανόνες τόσο το καλύτερο. Ευτυχώς, είναι συχνά δυνατό να καταστεί ένα σύνθετο δέντρο σε ένα μικρότερο σύνολο κανόνων.

Στο δέντρο απόφασης, κάθε αρχείο καταλήγει σε ακριβώς ένα φύλλο, έτσι κάθε αρχείο έχει μια οριστική ταξινόμηση. Ως εκ τούτου μετά από τη διαδικασία κανόνα-γενίκευσης, μπορούν να υπάρξουν κανόνες που δεν είναι αμοιβαία αποκλειστικοί και αρχεία που δεν καλύπτονται από οποιοδήποτε κανόνα. Συγκεκριμένα η επιλογή ενός κανόνα όταν περισσότεροι από αυτούς ισχύουν μπορεί να λύσει το πρώτο πρόβλημα. Το δεύτερο πρόβλημα απαιτεί την εισαγωγή μιας *κατηγορίας προεπιλογής* που ορίζεται σε οποιοδήποτε αρχείο το οποίο δεν καλύπτεται από οποιουσδήποτε από τους κανόνες. Χαρακτηριστικά, η κατηγορία που εμφανίζετε ποιο συχνά επιλέγεται και ως προεπιλογή.

Μόλις δημιουργήσει ένα σύνολο γενικευμένων κανόνων, ο αλγόριθμος ομαδοποιεί τους κανόνες για κάθε κατηγορία και αποβάλλει εκείνους που δεν φαίνονται να συμβάλλουν στην ακρίβεια του συνόλου κανόνων. Το τελικό αποτέλεσμα είναι ένας μικρός αριθμός εύκολα κατανοητών κανόνων. (Robert Elliot, 2001)

2.5.5 Δέντρα απόφασης στην πράξη

Τα δέντρα απόφασης μπορούν να εφαρμοστούν σε πολλές διαφορετικές καταστάσεις, όπως:

- Για να ερευνήσουν ένα μεγάλο σύνολο δεδομένων και να επιλέξει τις χρήσιμες μεταβλητές
- Για να προβλέψουν τις μελλοντικές καταστάσεις των σημαντικών μεταβλητών σε μια βιομηχανική διαδικασία

- Για να διαμορφώσουν τις κατευθυνόμενες συστάδες των πελατών για ένα σύστημα σύστασης

2.6 Τεχνητά νευρωνικά δίκτυα

Τα τεχνητά νευρωνικά δίκτυα είναι δημοφιλή επειδή έχουν ένα αποδεδειγμένο αρχείο διαδρομής σε πολλές εφαρμογές εξόρυξης δεδομένων και εφαρμογές απόφασης-υποστήριξης. Τα νευρωνικά δίκτυα είναι συνήθως μια κατηγορία ισχυρών γενικής χρήσης εργαλείων που εφαρμόζονται εύκολα στην πρόβλεψη, την ταξινόμηση, και τη συγκέντρωση. Επίσης έχουν εφαρμοστεί από μια ευρεία σειρά βιομηχανιών, από την πρόβλεψη χρονικών σειρών στο οικονομικό περιβάλλον στη διάγνωση φυσικών καταστάσεων, από τον προσδιορισμό των συστάδων των πολύτιμων πελατών, στους ελέγχους και στην πρόβλεψη των ποσοστών αποτυχίας μηχανών.

Τα βιολογικά νευρωνικά δίκτυα είναι το ισχυρότερο είδος νευρωνικών δικτύων. Συγκεκριμένα ο ανθρώπινος εγκέφαλος καθιστά πιθανό για τους ανθρώπους να γενικεύσουν εμπειρικά, οι υπολογιστές, αφ' ετέρου, υπερέχουν συνήθως στο να ακολουθούν ρητές οδηγίες επανειλημμένως. Η έφεση των νευρωνικών δικτύων είναι ότι γεφυρώνουν αυτό το χάσμα με τη διαμόρφωση, ενός προτύπου σε έναν ψηφιακό υπολογιστή, των νευρωνικών συνδέσεων του ανθρώπινου εγκεφάλου. Κατ’ επέκταση όταν χρησιμοποιούνται σε καθορισμένες και αποσαφηνισμένες περιοχές, η δυνατότητά τους να γενικεύσουν και να μάθουν από τα στοιχεία μιμείται, υπό κάποια έννοια, τη δυνατότητά μας να μάθουμε από την εμπειρία. Αυτή η δυνατότητα είναι χρήσιμη για την εξόρυξη δεδομένων, και καθιστά επίσης τα νευρωνικά δίκτυα έναν συναρπαστικό τομέα για την έρευνα, που υπόσχεται νέα και καλύτερα αποτελέσματα στο μέλλον.

Ωστόσο υπάρχει ένα μειονέκτημα,. Τα αποτελέσματα της κατάρτισης (εκπαίδευσης) ενός νευρωνικού δικτύου είναι οι συντελεστές βάρους που διανέμονται σε όλο το δίκτυο. Αυτά οι συντελεστές βάρους δεν παρέχουν διορατικότητα *για το αν* η λύση ισχύει. Ίσως μια ημέρα, οι περίπλοκες τεχνικές για τα νευρωνικά δίκτυα μπορούν να βοηθήσουν και να παρέχουν κάποια εξήγηση. Στο μεταξύ, τα νευρωνικά δίκτυα προσεγγίζονται καλύτερα ως μαύρα κουτιά με εσωτερικά έργα τόσο μυστήρια όσο τα έργα των εγκεφάλων μας (Berry and Linoff, 2004).

Οι απαντήσεις που παράγονται από τα νευρωνικά δίκτυα είναι συχνά σωστές. Έχουν αξία για την επιχείρηση και σε πολλές περιπτώσεις, ένα σημαντικό χαρακτηριστικό γνώρισμα από το να παρέχουν μια εξήγηση.

Τα νευρωνικά δίκτυα έχουν τη δυνατότητα να μαθαίνουν από τα παραδείγματα με τον ίδιο σχεδόν τρόπο που οι ανθρώπινοι κερδίζουν γνώση από την εμπειρία.

2.6.1 Νευρωνικά δίκτυα για την κατευθυνόμενη εξόρυξη δεδομένων

Η πιο κοινή χρήση των νευρωνικών δικτύων είναι η οικοδόμηση ενός προτύπου για την ταξινόμηση ή την πρόβλεψη. Τα βήματα σε αυτήν την διαδικασία είναι:

1. Προσδιορισμός των χαρακτηριστικών γνωρισμάτων εισαγωγής και εξαγωγής.
2. Μετασχηματισμός των εισαγωγών και των αποτελεσμάτων έτσι ώστε να είναι σε μια μικρή κλίμακα, (-1 έως 1).
3. Οργάνωση ενός δικτύου με μια κατάλληλη τοπολογία.
4. Εκπαίδευση του δικτύου σε ένα αντιπροσωπευτικό σύνολο δεδομένων
5. Χρησιμοποίηση συνόλων επικύρωσης έτσι ώστε να επιλεγεί το σύνολο των κριτηρίων που ελαχιστοποιούν το λάθος.
6. Αξιολόγηση του δικτύου χρησιμοποιώντας τη δοκιμή για να διαπιστωθεί πόσο καλά λειτουργεί.
7. Εφαρμογή του πρότυπου που παράγεται από το δίκτυο για να προβλέψει εκβάσεις που αφορούν τις άγνωστες εισαγωγές.

Ευτυχώς, το λογισμικό εξόρυξης δεδομένων εκτελεί τα περισσότερα από αυτά τα βήματα αυτόματα. Αν και η γνώση των εσωτερικών εργασιών δεν είναι απαραίτητη, υπάρχουν μερικά κλειδιά που επιτρέπουν την επιτυχή χρήση των δικτύων. Αρχικά όπως με όλα τα προβλεπτικά εργαλεία διαμόρφωσης, το σημαντικότερο ζήτημα είναι να επιλεγεί το σωστό σύνολο δεδομένων προς εκπαίδευση. Δεύτερο ζήτημα είναι τα στοιχεία να αντιπροσωπεύονται με τέτοιο τρόπο ώστε να μεγιστοποιείτε η δυνατότητα του δικτύου να αναγνωρίζει πρότυπα σε αυτό. Το τρίτο είναι να ερμηνεύει τα αποτελέσματα από το δίκτυο. Τέλος, η κατανόηση μερικών συγκεκριμένων λεπτομερειών όσον αφορά το πώς λειτουργούν, όπως η τοπολογία δικτύων και οι παράμετροι που ελέγχουν την κατάρτιση, μπορεί να βοηθήσει στην δημιουργία καλύτερων δικτύων εκτέλεσης.

Επίσης ένας από τους κινδύνους όπως και με οποιοδήποτε πρότυπο που χρησιμοποιείται για την πρόβλεψη ή την ταξινόμηση είναι το πρότυπο να χάνει την

χρησιμότητα του καθώς περνάει ο καιρός και τα νευρωνικά πρότυπα δικτύων δεν είναι εξαίρεση σε αυτόν τον κανόνα.

Τέλος ένα νευρωνικό δίκτυο είναι τόσο καλό όσο το σύνολο κατάρτισης (σύστημα εκπαίδευσης) που χρησιμοποιήθηκε για να το παράγει. Το πρότυπο είναι στατικό και πρέπει να ενημερώνετε ρητά με την προσθήκη των πιο πρόσφατων παραδειγμάτων στο σύνολο των δεδομένων προς εκπαίδευση όπως επίσης και στην επανεκπαίδευση του δικτύου (ή την εκπαίδευση ενός νέου δικτύου) προκειμένου να το κρατήσει ενημερωμένο και χρήσιμο. (Rygielski C. 2002)

2.6.2 Τι είναι ένα νευρωνικό δίκτυο

Τα νευρωνικά δίκτυα αποτελούνται από βασικές μονάδες που μιμούνται, με έναν απλουστευμένο τρόπο, τη συμπεριφορά των βιολογικών νευρώνων που βρίσκονται στη φύση. Η βασική ιδέα είναι ότι κάθε νευρωνική μονάδα, έχει πολλές εισαγωγές που η κάθε μονάδα συνδυάζεται σε μια ενιαία αξία παραγωγής. Στους εγκεφάλους, αυτές οι μονάδες μπορούν να συνδεθούν με τα εξειδικευμένα νεύρα. Εν τούτοις οι υπολογιστές, είναι λίγο απλούστεροι και οι μονάδες συνδέονται απλά, έτσι τα αποτελέσματα από μερικές μονάδες χρησιμοποιούνται ως εισαγωγές σε άλλες.

Το δίκτυο τροφοδοσίας προς τα εμπρός είναι ο απλούστερος και πιο χρήσιμος τύπος δικτύου για την κατευθυνόμενη διαμόρφωση.

2.6.3 Νευρωνικά δίκτυα με προς τα εμπρός τροφοδοσία (feed-forward networks)

Ένα νευρωνικό δίκτυο με προς τα εμπρός τροφοδοσία υπολογίζει τις παραγόμενες τιμές από τις τιμές εισαγωγής. Συγκεκριμένα η τοπολογία ή η δομή, αυτού του δικτύου είναι χαρακτηριστικές των δικτύων που χρησιμοποιούνται για την πρόβλεψη και την ταξινόμηση. Οι μονάδες οργανώνονται σε τρία στρώματα.

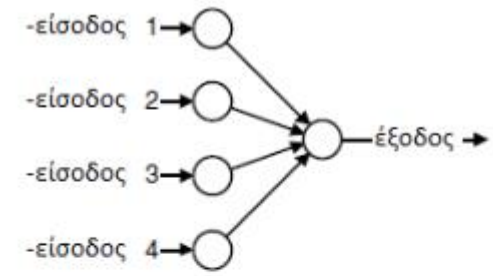
Το στρώμα εισαγωγής. Κάθε μονάδα στο στρώμα εισαγωγής συνδέεται με ακριβώς έναν τομέα πηγής, ο οποίος έχει χαρτογραφηθεί χαρακτηριστικά στη σειρά -1 έως 1. Έπειτα κάθε μονάδα του στρώματος εισαγωγής αντιγράφει την αξία εισαγωγής της στην παραγωγή της.

Το επόμενο στρώμα καλείται *κρυμμένο στρώμα* επειδή δεν συνδέεται ούτε με τις εισαγωγές ούτε με την παράγωγη του δικτύου. Συγκεκριμένα κάθε μονάδα στο κρυμμένο στρώμα συνδέεται τυπικά με όλες τις μονάδες στο στρώμα εισαγωγής. Κατ' επέκταση δεδομένου ότι αυτό το δίκτυο περιέχει τις τυποποιημένες μονάδες, οι μονάδες στο κρυμμένο στρώμα υπολογίζουν την παράγωγή τους με τον πολλαπλασιασμό της αξίας κάθε εισαγωγής

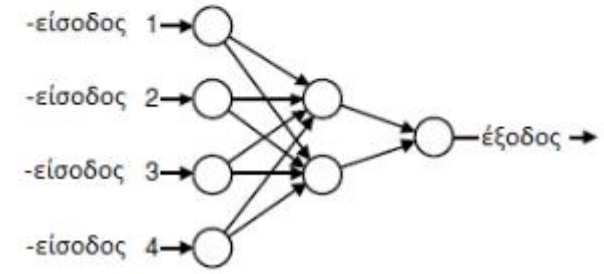
με το αντίστοιχο ‘‘βάρος’’ της, προσθέτοντας της, και εφαρμόζοντας την λειτουργία μεταφοράς. Επίσης ένα νευρωνικό δίκτυο μπορεί να έχει έναν οποιοδήποτε αριθμό κρυμμένων στρωμάτων, αλλά γενικά, ένα κρυμμένο στρώμα είναι ικανοποιητικό. Κατ’ επέκταση όσο ευρύτερο το στρώμα (δηλαδή όσο περισσότερες μονάδες περιέχει) τόσο μεγαλύτερη η ικανότητα του δικτύου να αναγνωρίζει τα σχέδια. Ωστόσο αυτή η ικανότητα έχει ένα μειονέκτημα, επειδή το νευρωνικό δίκτυο μπορεί να απομνημονεύσει πρότυπα του ενός στα παραδείγματα εκπαίδευσης. *Το δίκτυο πρέπει να γενικεύει στο σύνολο κατάρτισης, και όχι να το απομνημονεύει*. Για να το επιτύχει αυτό, το κρυμμένο στρώμα δεν πρέπει να είναι πάρα πολύ ευρύ.

Το τελευταίο στρώμα είναι το *στρώμα παραγωγής* επειδή συνδέεται με την παραγωγή του νευρωνικού δικτύου και συνδέεται πλήρως με όλες τις μονάδες στο κρυμμένο στρώμα. Τις περισσότερες φορές, το νευρωνικό δίκτυο χρησιμοποιείται για να υπολογίσει μια μοναδική αξία, έτσι υπάρχει μόνο μια μονάδα στο στρώμα παραγωγής και την αξία. Επίσης σε μερικές εφαρμογές, το στρώμα παραγωγής χρησιμοποιεί μια απλή γραμμική λειτουργία μεταφοράς, έτσι η παραγωγή είναι ένας σταθμισμένος γραμμικός συνδυασμός εισαγωγών. Κατ’ επέκταση αυτό εξαλείφει την ανάγκη να χαρτογραφηθούν τα αποτελέσματα. Επίσης είναι δυνατό να υπάρξουν περισσότερες από μια μονάδες στο στρώμα παραγωγής.

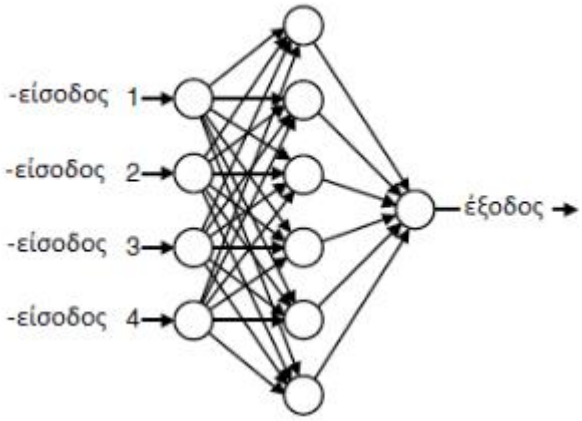
Υπάρχουν διάφορες παραλλαγές στην τοπολογία των νευρωνικών δικτύων με προς τα εμπρός τροφοδοσία. Μερικές φορές, τα στρώματα εισαγωγής συνδέονται άμεσα με το στρώμα παραγωγής. Σε αυτήν την περίπτωση, το δίκτυο έχει δύο συστατικά. Αυτές οι άμεσες συνδέσεις συμπεριφέρονται όπως μια τυποποιημένη οπισθοδρόμηση (γραμμικές ή λογιστικές, ανάλογα με τη λειτουργία ενεργοποίησης στο στρώμα παραγωγής). Αυτό είναι χρήσιμο στην δημιουργία περισσότερο τυποποιημένων στατιστικών πρότυπων. Έπειτα το κρυμμένο στρώμα ενεργεί ως ρυθμιστής του στατιστικού πρότυπου.



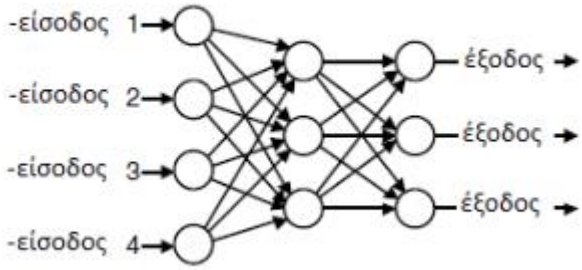
Αυτό το απλό νευρωνικό δίκτυο παίρνει τέσσερις εισόδους και παράγει μια έξοδο (παραγόμενη αξία) Το αποτέλεσμα της εκπαίδευσης ενός τέτοιου δικτύου είναι ισάξιο της στατιστικής τεχνικής που ονομάζεται λογιστική οπισθοδρόμηση.



Αυτό το δίκτυο έχει ένα μεσαίο στρώμα που ονομάζεται κρυμμένο στρώμα το οποίο κάνει το δίκτυο πιο ισχυρό επιτρέποντας το να αναγνωρίζει περισσότερα πρότυπα



Αυξάνοντας τον αριθμό των κρυμμένων στρωμάτων το δίκτυο γίνετε πιο ισχυρό όμως αυξάνετε ο κίνδυνος της υπερπροσαρμογής (overfitting). Συνήθως ένα κρυμμένο στρώμα είναι αρκετό.



Ένα νευρωνικό δίκτυο μπορεί να παράγει πολλαπλές αξίες παραγωγής.

Σχήμα 2.2 Τα νευρωνικά δίκτυα με προς τα εμπρός τροφοδοσία λαμβάνουν εισόδους στο ένα άκρο και τις μετατρέπουν σε εκροές (Berry and Linoff, 2004)

2.6.4 Επιλογή του συνόλου δεδομένων προς εκπαίδευση

Το σύνολο δεδομένων προς εκπαίδευση (training set) αποτελείται από αρχεία των οποίων οι τιμές πρόβλεψης ή ταξινόμησης είναι ήδη γνωστές. Η επιλογή ενός καλού συνόλου δεδομένων προς εκπαίδευση είναι κρίσιμη για όλη τη διαμόρφωση της εξόρυξης δεδομένων. Για παράδειγμα ένα φτωχό σύνολο καταδικάζει το δίκτυο, ανεξάρτητα από οποιαδήποτε άλλη εργασία που συμβάλει στη δημιουργία του. Ευτυχώς, υπάρχουν μόνο μερικά πράγματα που εξετάζουν την επιλογή ενός σωστού συστήματος εκπαίδευσης. (Berry and Linoff, 2004).

α) *Κάλυψη των τιμών για όλα τα χαρακτηριστικά γνωρίσματα.*

Το σημαντικότερο αυτών των εκτιμήσεων είναι ότι οι καθορισμένες ανάγκες κατάρτισης πρέπει να καλύπτουν την πλήρη σειρά των τιμών για όλα τα χαρακτηριστικά γνωρίσματα που το δίκτυο μπορεί να αντιμετωπίσει, συμπεριλαμβανομένης και της παραγωγής.

β) *Αριθμός χαρακτηριστικών γνωρισμάτων.*

Ο αριθμός των χαρακτηριστικών γνωρισμάτων εισόδου έχει επιπτώσεις στα νευρωνικά δίκτυα με δύο τρόπους. Κατ' αρχάς, όσο περισσότερα χαρακτηριστικά γνωρίσματα χρησιμοποιούνται ως εισαγωγές στο δίκτυο, τόσο μεγαλύτερο πρέπει να είναι το δίκτυο, αυξάνοντας τον κίνδυνο το πρότυπο να είναι υπερβολικά σύνθετο, με την κατοχή παρά πολλών παραμέτρων σχετικά με τον αριθμό παρατηρήσεων και αυξάνοντας το μέγεθος του συνόλου δεδομένων προς εκπαίδευση. Δεύτερον, όσο περισσότερα τα χαρακτηριστικά γνωρίσματα, τόσο περισσότερο χρόνο παίρνουν τα δίκτυα στην σύγκλιση ενός συνόλου σταθμών. Κατ' επέκταση με πάρα πολλά χαρακτηριστικά γνωρίσματα, τα σταθμά είναι λιγότερο πιθανό να είναι βέλτιστα.

γ) Μέγεθος του συνόλου δεδομένων προς εκπαίδευση

Όσο περισσότερα είναι τα χαρακτηριστικά γνωρίσματα που βρίσκονται στο δίκτυο, τόσα είναι και τα παραδείγματα κατάρτισης που απαιτούνται για μια καλή κάλυψη των σχεδίων στα στοιχεία. Δυστυχώς, δεν υπάρχει κανένας απλός κανόνας για να εκφραστεί μια σχέση μεταξύ του αριθμού χαρακτηριστικών γνωρισμάτων και του μεγέθους του συνόλου δεδομένων προς εκπαίδευση.

δ) Αριθμός αποτελεσμάτων

Στα περισσότερα παραδείγματα κατάρτισης, υπάρχουν πολλές περισσότερες εισαγωγές από ότι εξαγωγές αποτελεσμάτων, έτσι η καλή κάλυψη των εισαγωγών οδηγεί στην καλή κάλυψη των αποτελεσμάτων. Εντούτοις, είναι πολύ σημαντικό να υπάρχουν πολλά παραδείγματα για όλες τις πιθανές τιμές παραγωγής από το δίκτυο. Επιπλέον, ο αριθμός παραδειγμάτων

κατάρτισης για κάθε πιθανή παραγωγή πρέπει να είναι σχεδόν ίδιος. Αυτό μπορεί να είναι κρίσιμο κατά την απόφαση της επιλογής του συνόλου κατάρτισης που θα χρησιμοποιηθεί.

2.7 Προσέγγιση κοντινότερου γείτονα

Η τεχνική της προσέγγισης του κοντινότερου γείτονα είναι βασισμένη στην έννοια της ομοιότητας. Τα βασισμένα στην μνήμη αποτελέσματα συλλογισμού (MBR) είναι βασισμένα σε ανάλογες καταστάσεις του παρελθόντος όπως για παράδειγμα την απόφαση ότι ένας νέος φίλος είναι Αυστραλός, βασιζόμενοι σε προηγούμενα παραδείγματα αυστραλιανής προφοράς. Συγκεκριμένα το συνεργάσιμο φιλτράρισμα προσθέτει περισσότερες πληροφορίες, χρησιμοποιώντας όχι μόνο τις ομοιότητες μεταξύ των γειτόνων, αλλά και τις προτιμήσεις τους. Ένα παράδειγμα συνεργατικού φιλτραρίσματος είναι η σύσταση ενός εστιατορίου.

Το κέντρο σε όλες αυτές τις τεχνικές είναι η ιδέα της ομοιότητας. Δηλαδή τι είναι αυτό που κάνει μια κατάσταση του παρελθόντος παρόμοια με μια νέα κατάσταση. Παρόλα αυτά μαζί με την εύρεση των παρόμοιων αρχείων του παρελθόντος, υπάρχει και η πρόκληση του συνδυασμού των πληροφοριών από τους γείτονες. Αυτές είναι οι δύο βασικές έννοιες της τεχνικής της προσέγγισης του κοντινότερου γείτονα. (Berry and Linoff, 2004)

2.7.1 Βασισμένα στη μνήμη αποτελέσματα συλλογισμού (Memory Based Reasoning)

Η ανθρώπινη δυνατότητα να αιτιολογεί μέσα από την εμπειρία εξαρτάται από τη δυνατότητα αναγνώρισης των κατάλληλων παραδειγμάτων του παρελθόντος. Ας πάρουμε για παράδειγμα ένα γιατρό που προσπαθεί να διαγνώσει ασθένειες, και έναν αναλυτή αξιώσεων που προσπαθεί να εντοπίσει ψευδείς ασφαλιστικές αξιώσεις, όλοι τους ακολουθούν μια παρόμοια διαδικασία. Ο καθένας πρώτα προσδιορίζει παρόμοιες περιπτώσεις από την εμπειρία του και έπειτα εφαρμόζει τι γνώση αυτών των περιπτώσεών έτσι ώστε να προσεγγίσει το πρόβλημα. Αυτή είναι η ουσία του βασισμένου στην μνήμη συλλογισμού. Ποιο συγκεκριμένα μια βάση δεδομένων γνωστών αρχείων αναζητάτε για να βρει τα αρχεία που είναι παρόμοια με το νέο αρχείο. Έπειτα αυτοί οι γείτονες χρησιμοποιούνται για την ταξινόμηση και την εκτίμηση. Οι εφαρμογές MBR εκτείνονται σε πολλές περιοχές:

Ανίχνευση απάτης. Νέες υποθέσεις απάτης είναι πιθανό να είναι παρόμοιες με είδη γνωστές περιπτώσεις. Το MBR μπορεί να τις εντοπίσει και να τις "μαρκάρει" για περαιτέρω έρευνα.

Πρόβλεψη ανταπόκρισης πελατών. Οι επόμενοι πελάτες που είναι πιθανό να ανταποκριθούν σε μια προσφορά είναι πιθανώς παρόμοιοι με προηγούμενους πελάτες που

έχουν ανταποκριθεί. Ως εκ τούτου το MBR μπορεί εύκολα να προσδιορίσει τους επόμενους πιθανούς πελάτες.

Ιατρική περίθαλψη. Η αποτελεσματικότερη θεραπεία για έναν δεδομένο ασθενή είναι πιθανώς η θεραπεία που οδήγησε στα καλύτερα αποτελέσματα άλλους παρόμοιους ασθενείς. Το MBR μπορεί να βρει την διαδικασία η οποία παράγει το καλύτερο αποτέλεσμα.

Ταξινόμηση των απαντήσεων. Οι ελεύθερες απαντήσεις, όπως εκείνες στη μορφή απογραφής για το επάγγελμα και τη βιομηχανία ή τα φύλλα παραπόνων που προέρχονται από τους πελάτες, πρέπει να ταξινομηθούν σε ένα σταθερό σύνολο κωδίκων. Το MBR μπορεί να επεξεργαστεί αυτά τα ελεύθερα κείμενα και να ορίσει κώδικες για αυτά.

Μια από τις δυνάμεις των MBR είναι η δυνατότητά του να χρησιμοποιεί τα στοιχεία "όπως είναι." Αντίθετα από άλλες τεχνικές εξόρυξης δεδομένων, δεν ενδιαφέρετε για την μορφή των αρχείων. Ποιό συγκεκριμένα ενδιαφέρετε μόνο για την ύπαρξη δύο διαδικασιών: Μια λειτουργία απόστασης ικανή να υπολογίζει την απόσταση μεταξύ οποιωνδήποτε δύο αρχείων και μιας λειτουργίας συνδυασμού ικανής να συνδυάζει αποτελέσματα που προκύπτουν από διάφορους γείτονες για να φθάσει σε μια απάντηση. Αυτές οι λειτουργίες καθορίζονται εύκολα για πολλά είδη αρχείων, συμπεριλαμβανομένων των αρχείων με σύνθετους ή ασυνήθιστους τύπους στοιχείων όπως οι γεωγραφικές θέσεις, οι εικόνες, και το ελεύθερο κείμενο που είναι συνήθως δύσκολο να τα χειριστούμε με άλλες τεχνικές ανάλυσης. Μια άλλη δύναμη των MBR είναι η δυνατότητά τους να προσαρμόζονται. Συγκεκριμένα μόνο από την ενσωμάτωση νέων στοιχείων στην ιστορική βάση δεδομένων καθιστά πιθανό για τα MBR να μάθουν για νέες κατηγορίες και για νέους ορισμούς από τα παλιά στοιχεία. Επίσης τα MBR παράγουν καλά αποτελέσματα χωρίς να χρειάζεται να αφιερωθεί μια μεγάλη περίοδος στην κατάρτιση (εκπαίδευση) ή στην μετατροπή των εισερχόμενων στοιχείων στη σωστή φόρμα.

Ωστόσο όλα αυτά τα πλεονεκτήματα έρχονται με ένα κόστος. Το MBR τείνει να απασχολεί υπερβολικά τους πόρους δεδομένων γιατί ένα μεγάλο ποσό ιστορικών στοιχείων πρέπει να είναι εύκολα διαθέσιμο για την εύρεση των γειτόνων. Επίσης η ταξινόμηση των νέων αρχείων μπορεί να απαιτήσει την επεξεργασία όλων των ιστορικών αρχείων για να εντοπίσει τους ποιο όμοιους γείτονες, μια πολύ πιο χρονοβόρα διαδικασία από το να εφαρμοστεί ένα ήδη εκπαιδευμένο νευρωνικό δίκτυο ή ένα ήδη χτισμένο δέντρο απόφασης. Υπάρχει επίσης η πρόκληση της εύρεσης των καλών λειτουργιών απόστασης και συνδυασμού, η οποίες απαιτούν συχνά μια δοκιμή λάθους και καλή διαίσθηση. (Robert Elliot, 2001)

2.7.2 Προκλήσεις MBR

Γενικά, η χρησιμοποίηση των MBR περιλαμβάνει διάφορες επιλογές:

1. Επιλογή ενός κατάλληλου συνόλου αρχείων κατάρτισης
2. Επιλογή του αποδοτικότερου τρόπου να αντιπροσωπευθούν τα αρχεία κατάρτισης
3.Επιλογή της λειτουργίας απόστασης, της λειτουργίας συνδυασμού, και του αριθμού γειτόνων.

Εξετάζετε κάθε μια από αυτές στη συνέχεια.

1) Επιλογή ενός ισορροπημένου συνόλου ιστορικών αρχείων

Το σύνολο κατάρτισης είναι ένα σύνολο ιστορικών αρχείων. Ως εκ τούτου πρέπει να παρέχει μια καλή κάλυψη του πληθυσμού έτσι ώστε οι κοντινότεροι γείτονες ενός άγνωστου αρχείου να μπορούν να είναι χρήσιμοι για προγνωστικούς λόγους. Επίσης ένα τυχαίο δείγμα μπορεί να μην παρέχει ικανοποιητική κάλυψη για όλες τις τιμές. Μερικές κατηγορίες είναι συχνότερες από άλλες και οι συχνότερες αυτές κατηγορίες εξουσιάζουν το τυχαίο δείγμα. Παραδείγματος χάριν, οι ψευδείς συναλλαγές είναι πολύ σπανιότερες από τις μη ψευδείς συναλλαγές, ή οι καρδιακές παθήσεις είναι πιο κοινές από τον καρκίνο του συκωτιού, και τα λοιπά.

Για να επιτύχει την ισορροπία, το σύνολο κατάρτισης πρέπει, εάν είναι δυνατόν, να περιέχει τους κατά προσέγγιση ίσους αριθμούς αρχείων που αντιπροσωπεύουν τις διαφορετικές κατηγορίες (Berry and Linoff, 2004).

2) Αντιπροσώπευση των στοιχείων κατάρτισης

Η απόδοση των MBR στην παραγωγή προβλέψεων εξαρτάται από τον τρόπο με τον οποίο αντιπροσωπεύεται το σύνολο δεδομένων προς εκπαίδευση. Ποιο συγκεκριμένα η προσέγγιση πλοκών διασποράς λειτουργεί για δύο ή τρεις μεταβλητές και έναν μικρό αριθμό αρχείων, αλλά δεν κάνει καλή ιεράρχηση της κλίμακας. Η απλούστερη μέθοδος για την προσέγγιση του κοντινότερου γείτονα απαιτεί τον καθορισμό της απόστασης από την άγνωστη περίπτωση σε κάθε ένα από τα αρχεία του συνόλου δεδομένων προς εκπαίδευση και την επιλογή των αρχείων του συνόλου με τις μικρότερες αποστάσεις. Δεδομένου ότι ο αριθμός αρχείων αυξάνεται, ο χρόνος που απαιτείται για να βρεθούν οι γείτονες ενός νέου αρχείου αυξάνεται γρήγορα. Αυτό ισχύει ιδιαίτερα εάν τα αρχεία αποθηκεύονται σε μια σχεσιακή βάση δεδομένων.

Οι επιδώσεις των σχεσιακών βάσεων δεδομένων είναι αρκετά καλές σήμερα. Ωστόσο η πρόκληση με τη βαθμολόγηση των στοιχείων για τα MBR είναι ότι κάθε περίπτωση που

σημειώνεται πρέπει να συγκριθεί σε κάθε περίπτωση στη βάση δεδομένων. Η σημείωση ενός ενιαίου νέου αρχείου δεν παίρνει πολύ χρόνο, ακόμα και όταν υπάρχουν εκατομμύρια αρχείων. Παρόλα αυτά, η σημείωση πολλών νέων αρχείων μπορεί να έχει κακή απόδοση.

Ένας άλλος τρόπος να κατασταθεί το MBR αποδοτικότερο είναι να μειωθεί ο αριθμός των αρχείων στο σύνολο κατάρτισης. Για παράδειγμα είναι πιθανό τα περισσότερα από τα αρχεία να είναι περιττά, δηλαδή να μην είναι πραγματικά απαραίτητα για λόγους ταξινόμησης. (Berry and Linoff, 1999)

2.7.3 Καθορισμός της λειτουργίας απόστασης, της συνδυαστικής λειτουργίας, και του αριθμού γειτόνων

Η λειτουργία απόστασης, η λειτουργία συνδυασμού, και ο αριθμός γειτόνων είναι τα βασικά συστατικά στην χρησιμοποίηση των MBR. Κατ' επέκταση το ίδιο σύνολο ιστορικών αρχείων μπορεί να αποδειχθεί πολύ χρήσιμο ή και καθόλου χρήσιμο για προγνωστικούς λόγους, ανάλογα με αυτά τα κριτήρια. Ευτυχώς, οι απλές λειτουργίες απόστασης και οι λειτουργίες συνδυασμού λειτουργούν συνήθως αρκετά καλά.

Εφαρμογή MBR

Τα σημαντικά βήματα για την εφαρμογή των βασισμένων στην μνήμη αποτελεσμάτων συλλογισμού είναι:

1. Επιλογή του συνόλου κατάρτισης
2. Καθορισμός της λειτουργίας απόστασης
3. Επιλογή του αριθμού κοντινότερων γειτόνων
4. Καθορισμός της λειτουργίας συνδυασμού

1.Επιλογή του συνόλου κατάρτισης

Πολλές φορές το σύνολο κατάρτισης δεν δημιουργείται ειδικά για την περίπτωση που εξετάζεται, με αποτέλεσμα η συχνότητα των κωδίκων στο σύνολο κατάρτισης να ποικίλει πολύ, μιμώντας τη γενική συχνότητα των κωδίκων γενικά. Αν και αυτά τα σύνολα κατάρτισης μπορούν να παράγουν καλά αποτελέσματα, ένα καλύτερα κατασκευασμένο σύνολο κατάρτισης με περισσότερα παραδείγματα των λιγότερο κοινών κωδίκων πιθανώς να μπορεί να αποδώσει ακόμα καλύτερα.

2.Επιλογή της λειτουργίας απόστασης

Το επόμενο βήμα είναι η επιλογή της λειτουργίας απόστασης. Για παράδειγμα υπάρχει η περίπτωση, μια λειτουργία απόστασης να υπάρχει ήδη, βάσει της έννοιας της *σχετικής ανατροφοδότησης (relevance feedback)* που μετρά την ομοιότητα δύο εγγράφων βασισμένη στις λέξεις που περιέχουν. Η σχετική ανατροφοδότηση, είχε ως σκοπό αρχικά να επιστρέφει έγγραφα παρόμοια με ένα δεδομένο έγγραφο, ως έναν τρόπο αναζητήσεις. Κατ' επέκταση τα έγγραφα που παρουσιάζουν την μεγαλύτερη ομοιότητα είναι οι γείτονες που χρησιμοποιούνται για το MBR.

3. Επιλογή της λειτουργίας συνδυασμού

Η επόμενη απόφαση είναι η λειτουργία συνδυασμού. Τα περισσότερα προβλήματα ταξινόμησης ψάχνουν την ενιαία καλύτερη λύση. Εντούτοις, σε κάποιες περιπτώσεις μπορεί να υπάρχουν πολλαπλάσιοι κώδικες, ακόμη και από την ίδια κατηγορία. Η δυνατότητα προσαρμογής του MBR σε αυτό το πρόβλημα δίνει έμφαση στην ευελιξία του.

4.**Επιλογή του αριθμού γειτόνων**

Καλύτερα αποτελέσματα προέρχονται από τη χρησιμοποίηση περισσότερων γειτόνων. Εντούτοις το πιο χαρακτηριστικό πρόβλημα είναι να οριστεί μόνο μια ενιαία κατηγορία ή ένας κώδικας και λιγότεροι γείτονες, το οποίο θα ήταν αρκετό για να υπάρξουν καλά αποτελέσματα.

Μέτρηση της απόστασης

Ας πούμε για παράδειγμα πως θέλουμε να μάθουμε τον καιρό που θα έχει μια μικρή πόλη. Εάν για παράδειγμα έχουμε μια εφημερίδα που απαριθμεί τις καιρικές συνθήκες για μεγάλες πόλεις, τότε χαρακτηριστικά θα προσπαθήσουμε να βρούμε τον καιρό για τις πόλεις που βρίσκονται κοντά στη μικρή πόλη. Αυτό θα γίνει είτε εξετάζοντας την πιο κοντινή πόλη και παίρνοντας ακριβώς τον καιρό της, ή κάνοντας κάποιο είδος συνδυασμού των προβλέψεων, των τριών πιο κοντινών πόλεων. Αυτό είναι ένα παράδειγμα MBR για να βρεθεί μια καιρική πρόβλεψη. Επομένως η λειτουργία απόστασης που χρησιμοποιείται είναι η γεωγραφική απόσταση μεταξύ των δύο θέσεων.

Τι είναι μια λειτουργία απόστασης

Η απόσταση είναι ο τρόπος με τον οποίο το MBR μετρά την ομοιότητα . Για οποιαδήποτε αληθινή μετρική απόσταση, η απόσταση από το σημείο A στο σημείο B, που δείχνεται από το d (A,B), έχει τέσσερις βασικές ιδιότητες:

1.**Καλα καθορισμένο.** Η απόσταση μεταξύ δύο σημείων καθορίζεται πάντα και είναι ένας μη αρνητικός πραγματικός αριθμός, .

2.**Ταυτότητα.** Η απόσταση από ένα σημείο έως τον εαυτό του είναι πάντα μηδέν.

3.**Αντιμεταθετικοτιτα.** Η κατεύθυνση δεν κάνει διαφορά, έτσι η απόσταση από το A στο B είναι η ίδια με την απόσταση από το B στο A..

4.Ανισότητα τριγώνων. Η επίσκεψη ενός ενδιάμεσου τρίτου σημείου Γ
στον δρόμο από το A στο B δεν μικραίνει ποτέ την απόσταση.

Για το MBR, τα σημεία είναι τα πραγματικά αρχεία σε μια βάση δεδομένων. Επομένως αυτός ο επίσημος καθορισμός της απόστασης είναι η βάση για την μέτρηση της ομοιότητας.

Η λειτουργία συνδυασμού: Ερώτηση των γειτόνων για την απάντηση

Η λειτουργία απόστασης χρησιμοποιείται για να καθορίσει ποια αρχεία περιλαμβάνονται στη “γειτονιά”. Υπάρχουν διαφορετικοί τρόποι ώστε να συνδυαστούν τα στοιχεία που συγκεντρώνονται από τους γείτονες και να κάνουν μια πρόβλεψη.

Η βασική προσέγγιση: Δημοκρατία

Μια κοινή λειτουργία συνδυασμού είναι αυτή στην οποία οι K κοντινότεροί γείτονες ψηφίζουν για μια απάντηση, όπως η "Δημοκρατία", στην εξόρυξη δεδομένων. Όταν το MBR χρησιμοποιείται για την ταξινόμηση, κάθε γείτονας χρησιμοποιεί την ψήφο του για την κατηγορία στην οποία ανήκει. Επομένως το ποσοστό των ψηφοφοριών για κάθε κατηγορία είναι μια εκτίμηση της πιθανότητας ότι το νέο αρχείο ανήκει στην αντίστοιχη κατηγορία. Κατ’ επέκταση όταν ο στόχος είναι να οριστεί μια ενιαία κατηγορία, είναι απλά αυτή με τις περισσότερες ψήφους.

2.7.4 Συνεργατικό φιλτράρισμα: Μια προσέγγιση κοντινότερων γειτόνων στην υποβολή συστάσεων

Ένα ανθρώπινο παράδειγμα του συνεργατικού φιλτραρίσματος είναι η σύσταση από έμπιστους φίλους που θα αναγκάσει κάποιον να δοκιμάσει κάτι που κάτω από άλλες συνθήκες δεν θα το έκανε.

Ποιο συγκεκριμένα το συνεργάσιμο φιλτράρισμα είναι μια παραλλαγή του βασισμένου στην μνήμη συλλογισμού που ταιριάζει ιδιαίτερα καλά στην εφαρμογή παροχής των εξατομικευμένων συστάσεων. Ένα σύστημα συνεργατικού φιλτραρίσματος αρχίζει με ένα ιστορικό των προτιμήσεων των ανθρώπων. Έπειτα η λειτουργία απόστασης καθορίζει την ομοιότητα βασισμένη στην επικάλυψη των ανθρώπινων προτιμήσεων, ως εκ τούτου οι άνθρωποι που έχουν ίδιες προτιμήσεις είναι 'γείτονες'. Επιπλέον, οι ψηφοφορίες σταθμίζονται από τις αποστάσεις, έτσι οι ψήφοι των πιο κοντινών γειτόνων μετράν περισσότερο για τη σύσταση. Με άλλα λόγια, είναι μια τεχνική εύρεσης για τη μουσική, τα βιβλία, το κρασί, ή οτιδήποτε άλλο που μπορεί να αρμόζουν στις υπάρχουσες προτιμήσεις ενός ιδιαίτερου προσώπου με τη χρησιμοποίηση των κρίσεων μιας όμοιας ομάδας που επιλέγεται για τις παρόμοιες προτιμήσεις τους. Αυτή η προσέγγιση καλείται επίσης *κοινωνικό φιλτράρισμα πληροφοριών*.

Το συνεργατικό φιλτράρισμα αυτοματοποιεί τη προφορική διαδικασία μέσα από την οποία μαθαίνουμε τις προτιμήσεις. Παρόλα αυτά η γνώση ότι πολλοί άνθρωποι προτίμησαν κάτι δεν είναι αρκετή, είναι εξίσου σημαντικό το ποιος το προτίμησε και αυτό γιατί ο καθένας εκτιμάει κάποιες συστάσεις περισσότερο από άλλες.

Σύμφωνα με τους Berry και Linoff (2004) η προετοιμασία των συστάσεων για έναν νέο πελάτη που χρησιμοποιεί ένα αυτοματοποιημένο σύστημα συνεργατικού φιλτραρίσματος έχει τρία βήματα:

1. Την δημιουργία ενός σχεδιαγράμματος πελατών βάζοντας τον νέο πελάτη να εκτιμήσει μια επιλογή στοιχείων όπως τους κινηματογράφους, τα τραγούδια, ή τα εστιατόρια.

2. Σύγκριση του σχεδιαγράμματος του νέου πελάτη με τα σχεδιαγράμματα άλλων πελατών που χρησιμοποιούν κάποιο μέτρο ομοιότητας.

3. Χρησιμοποίηση κάποιου συνδυασμού των εκτιμήσεων των πελατών με παρόμοια σχεδιαγράμματα για να προβλέψει την εκτίμηση που θα έδινε ένας νέος πελάτης σε στοιχεία που δεν έχει εκτιμήσει ακόμα.

2.8 Συσταδοποίηση (Clustering)

Η συσταδοποίηση είναι η μέθοδος με την οποία τα αρχεία συγκεντρώνονται σε ομάδες. Συνήθως αυτό γίνεται για να δώσει στον τελικό χρήστη μια υψηλού επιπέδου άποψη για το τι συμβαίνει στη βάση δεδομένων. Η συσταδοποίηση χρησιμοποιείται μερικές φορές για να σημάνει την κατάτμηση - που οι περισσότεροι άνθρωποι του μάρκετινγκ θα πουν ότι είναι χρήσιμη για την επιχείρηση. Δύο από αυτά τα προγράμματα συσταδοποίησης είναι το σύστημα PRIZM™ από την εταιρία Claritas και το MicroVision™ από την εταιρία Equifax. Αυτές οι επιχειρήσεις έχουν ομαδοποιήσει τον πληθυσμό από τις δημογραφικές πληροφορίες σε τμήματα που θεωρούν ότι είναι χρήσιμα για το άμεσο μάρκετινγκ και τις πωλήσεις. Για να χτιστούν αυτές οι ομάδες χρησιμοποιούν πληροφορίες όπως το εισόδημα, η ηλικία, το επάγγελμα, την κατοικία και τη φυλή που έχουν συλλέξει από την απογραφή. Κατόπιν ορίζουν "παρωνύμια" στις συστάδες. Μερικά παραδείγματα παρουσιάζονται στον πίνακα 2.1.

Όνομα	Εισόδημα	Ηλικία	Μόρφωση	Πρόγραμμα
Blue Blood Estates	Wealthy	35-54	College	Claritas Prizm™
Shotguns and Pickups	Middle	35-64	High School	Claritas Prizm™
Southside City	Poor	Mix	Grade School	ClaritasPrizm™
Living Off the Land	Middle-Poor	SchoolAgeFamilies	Low	Equifax MicroVision™
University USA	Very low	Young - Mix	Medium to High	Equifax MicroVision™
Sunset Years	Medium	Seniors	Medium	Equifax MicroVision™

Πίνακας 2.1 *μερικές εμπορικά διαθέσιμες ετικέτες συστάδων*

Αυτές οι πληροφορίες συσταδοποίησης χρησιμοποιούνται έπειτα από τον τελικό χρήστη για να τοποθετήσει τους πελάτες στη βάση δεδομένων. Μόλις γίνει αυτό, ο χρήστης μπορεί να έχει άμεσα μια υψηλού επιπέδου άποψη για το τι συμβαίνει μέσα στη συστάδα. Μόλις εργαστεί με αυτούς τους κώδικες για κάποιο χρόνο αρχίζουν να χτίζονται διαισθήσεις για το πώς αυτές οι διαφορετικές συστάδες πελατών θα αντιδράσουν στις προσφορές

μάρκετινγκ στην επιχείρησή του. Παραδείγματος χάριν μερικές από αυτές τις συστάδες μπορούν να αφορούν την επιχείρησή του και μερικές από αυτές όχι. Αλλά δεδομένου ότι ο ανταγωνισμός μπορεί να χρησιμοποιεί αυτές τις ίδιες συστάδες για να κτίσει τις επιχειρησιακές προσφορές και προσφορές μάρκετινγκ είναι σημαντικό να γνωρίζει το πώς η βάση πελατών συμπεριφέρεται όσον αφορά αυτές τις συστάδες. (Alex Berson, Stephen Smith, Kurt Thearling , 1999)

Βρίσκοντας αυτά που δεν ταιριάζουν - συσταδοποίηση για ακραίες τιμές

Μερικές φορές η συσταδοποίηση εκτελείται όχι για να διατηρήσει τόσο πολύ μαζί τα αρχεία όσο για να καταστήσει ευκολότερο να βρεθεί το πότε ένα αρχείο διαφέρει από τα υπόλοιπα. Παραδείγματος χάριν: οι περισσότεροι διανομείς κρασιού που πωλούν φτηνό κρασί στο Μισούρι και στέλνουν ένα ορισμένο όγκο προϊόντων, έχουν ένα ορισμένο επίπεδο κέρδους. Υπάρχει μια ομάδα καταστημάτων που μπορεί να διαμορφωθεί με αυτά τα χαρακτηριστικά. Ένα κατάστημα ξεχωρίζει, εντούτοις, έχοντας σημαντικά χαμηλότερο κέρδος. Σε μια πιο προσεκτική εξέταση βρίσκουμε ότι ο διανομέας παρέδιδε το προϊόν αλλά δεν συνέλλεγε την πληρωμή από έναν από τους πελάτες του. Εκπτώσεις στα ανδρικά κοστούμια γίνονται σε όλους τους κλάδους ενός καταστήματος στη νότια Καλιφόρνια. Όλα τα καταστήματα με αυτά τα χαρακτηριστικά έχουν δει τουλάχιστον μια αύξηση 100% στα έσοδα από την έναρξη των εκπτώσεων εκτός από ένα. Τελικά βρέθηκε ότι αυτό το κατάστημα, αντίθετα από τους άλλους, είχε διαφημιστεί στο ραδιόφωνο παρά στην τηλεόραση.

2.8.1 Συσταδοποίηση και η τεχνική κοντινότερου γείτονα

Πώς η συσταδοποίηση είναι σαν την τεχνική κοντινότερου γείτονα;

Ο αλγόριθμος κοντινότερου γείτονα (nearest neighbor algorithm) είναι βασικά ένας καθορισμός της συσταδοποίησης υπό την έννοια ότι και οι δύο χρησιμοποιούν την απόσταση σε κάποιο διάστημα χαρακτηριστικών γνωρισμάτων για να δημιουργήσουν είτε τη δομή στα στοιχεία είτε τις προβλέψεις. Ο αλγόριθμος κοντινότερου γείτονα είναι μία εξέλιξη δεδομένου ότι μέρος του αλγορίθμου είναι συνήθως ένας τρόπος αυτόματης στάθμισης της σπουδαιότητας των προβλεπτών και πως θα μετρηθεί η απόσταση μέσα στο καθορισμένο διάστημα. Η συσταδοποίηση είναι μια ειδική περίπτωση όπου η σημασία κάθε προβλεπτή θεωρείται ισοδύναμη.

Πώς μπορείς να βάλεις τη συσταδοποίηση και τον κοντινότερο γείτονα να εργαστούν μαζί για την πρόβλεψη

Για να δούμε τη πρόβλεψη της συσταδοποίησης και του κοντινότερου γείτονα σε χρήση θα εργαστούμε με τη παρακάτω βάση δεδομένων και θα την εξετάσουμε με δύο τρόπους. Πρώτα προσπαθούμε να δημιουργήσουμε τις συστάδες μας - που εάν είναι χρήσιμες, θα μπορούσαμε να τις χρησιμοποιήσουμε εσωτερικά για να βοηθήσουν στην απλοποίηση και στη διευκρίνιση μεγάλης ποσότητας δεδομένων. Μετά προσπαθούμε να δημιουργήσουμε προβλέψεις βασισμένες στον κοντινότερο γείτονα.

Αρχικά ας ρίξουμε μια ματιά στα στοιχεία. Πώς θα ομαδοποιούσαμε τα στοιχεία στον πίνακα 2.2.

ID	Όνομα	Πρόβλεψη	Ηλικία	Έσοδα	Εισόδημα	Μάτια	Φύλο
1	Amy	Όχι	62	$0	Μέσο	Καστανά	Γ
2	Al	Όχι	53	$1,800	Μέσο	Πράσινα	Α
3	Betty	Όχι	47	$16,543	Υψηλό	Καστανά	Γ
4	Bob	Ναι	32	$45	Μέσο	Πράσινα	Α
5	Carla	Ναι	21	$2,300	Υψηλό	Μπλε	Γ
6	Carl	Όχι	27	$5,400	Υψηλό	Καστανά	Α
7	Donna	Ναι	50	$165	Χαμηλό	Μπλε	Γ
8	Don	Ναι	46	$0	Υψηλό	Μπλε	Α
9	Edna	Ναι	27	$500	Χαμηλό	Μπλε	Γ
10	Ed	Όχι	68	$1,200	Χαμηλό	Μπλε	Α

Πίνακας 2.2 *παράδειγμα μιας απλή βάση δεδομένων*

Εάν αυτοί ήταν φίλοι μας παρά πελάτες μας (ενδεχομένως μπορεί να είναι και τα δύο) και ήταν ελεύθεροι, μπορεί να τους ομαδοποιούσαμε βασισμένοι στη συμβατότητά του ενός με τον άλλον. Δημιουργώντας το δικό μας μίνι γραφείο γνωριμιών. Ένας πρακτικός τύπος θα συγκέντρωνε τη βάση δεδομένων ως εξής, επειδή σκέφτηκε ότι η συζυγική ευτυχία εξαρτάται συνήθως από την οικονομική συμβατότητα και θα δημιουργούνταν τρεις συστάδες όπως φαίνεται στον πίνακα 2.3.

ID	Όνομα	Πρόβλεψη	Ηλικία	Έσοδα	Εισόδημα	Μάτια	Φύλο

3	Betty	Όχι	47	$16,543	Υψηλό	Καστανά	Γ
5	Carla	Ναι	21	$2,300	Υψηλό	Μπλε	Γ

6	Carl	Όχι	27	$5,400	Υψηλό	Καστανά	Α
8	Don	Ναι	46	$0	Υψηλό	Μπλε	Α

1	Amy	Όχι	62	$0	Μέσο	Καστανά	Γ
2	Al	Όχι	53	$1,800	Μέσο	Πράσινα	Α
4	Bob	Ναι	32	$45	Μέσο	Πράσινα	Α

7	Donna	Ναι	50	$165	Χαμηλό	Μπλε	Γ
9	Edna	Ναι	27	$500	Χαμηλό	Μπλε	Γ
10	Ed	Όχι	68	$1,200	Χαμηλό	Μπλε	Α

Πίνακας 2.3 Μια απλή συσταδοποίηση της βάσης δεδομένων του παραδείγματος

Υπάρχει κάποιος άλλος "σωστός" τρόπος για να ομαδοποιήσουμε;

Εάν όμως ήμασταν περισσότερο «ρομαντικοί» μπορεί να εντοπίσαμε μερικές ασυμβατότητες μεταξύ του 46χρονου Don και της 21 ετών Carla (ακόμα κι αν και οι δύο έχουν πολύ καλά εισοδήματα). Αντ' αυτού να θεωρούσαμε την ηλικία και μερικά φυσικά χαρακτηριστικά σημαντικότερα στη δημιουργία των συστάδων των φίλων. Ένας άλλος τρόπος που θα μπορούσαμε να ομαδοποιήσουμε τους φίλους μας θα βασιζόταν στις ηλικίες και στο χρώμα των ματιών τους. Αυτό παρουσιάζεται στον πίνακα 2.4. Εδώ δημιουργούνται τρεις συστάδες όπου κάθε πρόσωπο στη συστάδα έχει σχεδόν ίδια ηλικία και έχει γίνει προσπάθεια να κρατηθούν εκείνοι που έχουν το ίδιο χρώμα ματιών μαζί στην ίδια συστάδα.

ID	Όνομα	Πρόβλεψη	Ηλικία	Έσοδα	Εισόδημα	Μάτια	Φύλο

5	Carla	Ναι	21	$2,300	Υψηλό	Μπλε	Γ
9	Edna	Ναι	27	$500	Χαμηλό	Μπλε	Γ
6	Carl	Όχι	27	$5,400	Υψηλό	Καστανά	Α
4	Bob	Ναι	32	$45	Μέσο	Πράσινα	Α

8	Don	Ναι	46	$0	Υψηλό	Μπλε	Α
7	Donna	Ναι	50	$165	Χαμηλό	Μπλε	Γ
10	Ed	Όχι	68	$1,200	Χαμηλό	Μπλε	Α

3	Betty	Όχι	47	$16,543	Υψηλό	Καστανά	Γ
2	Al	Όχι	53	$1,800	Μέσο	Πράσινα	Α
1	Amy	Όχι	62	$0	Μέσο	Καστανά	Γ

Πίνακας 2.4 Μία πιο "ρομαντική" ομαδοποίηση της βάσης δεδομένων του παραδείγματος

Δεν υπάρχει πιο "σωστός" τρόπος για συσταδοποίηση. Όπως βλέπουμε σε αυτό το παράδειγμα, αν και απλό, δίνει μερικές σημαντικές ερωτήσεις για τη συσταδοποίηση. Είναι δυνατό να πούμε εάν η πρώτη συσταδοποίηση που εκτελέστηκε παραπάνω (από την οικονομική θέση) ήταν καλύτερη ή χειρότερη από τη δεύτερη (κατά ηλικία και χρώμα ματιών); Πιθανότατα όχι, δεδομένου ότι οι συστάδες κατασκευάστηκαν για τον ιδιαίτερο σκοπό του να σημειώσουν τις ομοιότητες μεταξύ μερικών από τα δεδομένα και ότι η εικόνα της βάσης δεδομένων θα μπορούσε να είναι κάπως απλούστερη με τη χρησιμοποίηση της συσταδοποίησης. Αλλά ακόμη και οι διαφορές που δημιουργήθηκαν με τις δύο διαφορετικές περιπτώσεις συσταδοποίησης οδηγήθηκαν από ελαφρώς διαφορετικά κίνητρα (οικονομική κατάσταση εναντίον ρομαντισμού). Γενικά οι λόγοι για τη συσταδοποίηση δεν καθορίζονται εύκολα επειδή οι συστάδες χρησιμοποιούνται τις περισσότερες φορές για την εξερεύνηση και την περιληπτική παρουσίαση της πληροφορίας όσο και για την πρόβλεψη.

Πώς γίνονται οι ανταλλαγές κατά τον καθορισμό των αρχείων που χωρίζονται στις συστάδες;

Όπως παρατηρούμε, για το πρώτο παράδειγμα συσταδοποίησης υπήρξε ένας αρκετά απλός κανόνας με τον οποίο τα αρχεία θα μπορούσαν να χωριστούν σε συστάδες - συγκεκριμένα από το εισόδημα. Στο δεύτερο παράδειγμα συσταδοποίησης υπήρξαν λιγότερο σαφείς διαχωριστικές γραμμές δεδομένου ότι δύο προάγγελοι χρησιμοποιήθηκαν για να διαμορφώσουν τις συστάδες (ηλικία και χρώμα ματιών). Κατά συνέπεια η πρώτη συστάδα εξουσιάζεται από τους νέους με τα κάπως μικτά χρώματα ματιών ενώ οι τελευταίες δύο συστάδες έχουν ένα μίγμα των ηλικιωμένων όπου το χρώμα ματιών έχει χρησιμοποιηθεί για να τους χωρίσει (η δεύτερη συστάδα αποτελείται εξ ολοκλήρου από ανθρώπους με μπλε μάτια). Σε αυτήν την περίπτωση αυτές οι ανταλλαγές έγιναν αυθαίρετα αλλά κατά τη συσταδοποίηση πολύ μεγαλύτερου αριθμού αρχείων αυτές οι ανταλλαγές καθορίζονται ρητά από τον αλγόριθμο συσταδοποίησης.

Η συσταδοποίηση είναι ο ευτυχές διάμεσος μεταξύ των ομοιογενών συστάδων και του μικρότερου αριθμού συστάδων.

Στην καλύτερη δυνατή περίπτωση οι συστάδες θα χτίζονταν εκεί όπου όλα τα αρχεία μέσα στη συστάδα είχαν τις ίδιες τιμές για τους ιδιαίτερους προάγγελους που συγκεντρώνονταν. Αυτό θα ήταν το βέλτιστο στη δημιουργία μιας υψηλού επιπέδου άποψης αφού η γνώση των τιμών των προβλεπτών για οποιοδήποτε μέλος της συστάδας θα σήμαινε τη γνώση για τις τιμές για κάθε μέλος της συστάδας ανεξάρτητα από το πόσο μεγάλη ήταν. Η δημιουργία των ομοιογενών συστάδων όπου όλες οι τιμές για τους προάγγελους είναι ίδιες είναι δύσκολη να γίνει, όταν υπάρχουν πολλοί προάγγελοι ή/και οι προάγγελοι έχουν πολλές διαφορετικές τιμές (υψηλός αριθμός στοιχείων συνόλου).

Είναι δυνατό να εγγυηθεί ότι οι ομοιογενείς συστάδες δημιουργούνται με το να χωρίσουν οποιαδήποτε συστάδα που είναι ανομοιογενής σε μικρότερες συστάδες που είναι ομοιογενείς. Εν τούτοις, αυτό σημαίνει συνήθως τη δημιουργία συστάδων με μόνο ένα αρχείο που συνήθως υπερνικά τον αρχικό σκοπό της συσταδοποίησης. Παραδείγματος χάριν στην παραπάνω βάση δεδομένων μας με τα 10 αρχεία, 10 τέλεια ομοιογενείς συστάδες θα μπορούσαν να διαμορφωθούν με 1 αρχείο η κάθε μία, αλλά δεν θα είχε σημειωθεί αρκετή πρόοδος για να καταστήσει την αρχική βάση δεδομένων πιο κατανοητή.

Ο δεύτερος σημαντικός περιορισμός στη συσταδοποίηση είναι ότι διαμορφώνεται ένας σημαντικός αριθμός συστάδων. Όπου πάλι, η σημαντικότητα καθορίζεται από το χρήστη αλλά είναι δύσκολο να ποσολογήσει πέρα από αυτή εκτός από το για να πει ότι μόνο μια συστάδα δεν γίνεται αποδεκτή (μεγάλη γενίκευση) και ότι δεν γίνονται αποδεκτά τόσες συστάδες και αρχικά αρχεία. Πολλοί αλγόριθμοι συσταδοποίησης είτε αφήνουν το χρήστη να επιλέξει τον αριθμό συστάδων που θα επιθυμούσε να δει να δημιουργείτε από τη βάση δεδομένων ή παρέχουν στο χρήστη μια «διακλάδωση» από την οποία μπορούν να δημιουργηθούν μικρότεροι ή μεγαλύτεροι αριθμοί συστάδων, αμφίδρομα αφότου έχει εκτελεστεί η συσταδοποίηση.

Ποια είναι η διαφορά μεταξύ της συσταδοποίησης και της πρόβλεψης του κοντινότερου γείτονα;

Η κύρια διάκριση μεταξύ της συσταδοποίησης και της τεχνικής κοντινότερου γείτονα είναι ότι η συσταδοποίηση είναι μία ανεπίβλεπτη τεχνική εκμάθησης (unsupervised learning technique) και ο κοντινότερος γείτονας χρησιμοποιείται γενικά για την πρόβλεψη ή για μια εποπτευομένη τεχνική εκμάθησης (supervised learning technique). Οι ανεπίβλεπτες τεχνικές εκμάθησης είναι ανεπίβλεπτες υπό την έννοια ότι όταν οργανώνονται, δεν υπάρχει ιδιαίτερος λόγος για τη δημιουργία των προτύπων όπως υπάρχει για τις εποπτευόμενες τεχνικές εκμάθησης που προσπαθούν να εκτελέσουν πρόβλεψη. Στην πρόβλεψη, τα σχέδια που βρίσκονται στη βάση δεδομένων και παρουσιάζονται στο πρότυπο είναι πάντα τα σημαντικότερα σχέδια στη βάση δεδομένων για την εκτέλεση κάποιας ιδιαίτερης πρόβλεψης. Στη συσταδοποίηση δεν υπάρχει καμία ιδιαίτερη αίσθηση γιατί ορισμένα αρχεία είναι κοντά το ένα στο άλλο ή γιατί όλα περιέρχονται στην ίδια συστάδα. Μερικές από τις διαφορές μεταξύ της συσταδοποίησης και της πρόβλεψης κοντινότερου γείτονα μπορούν να συνοψιστούν στον πίνακα 2.5.

Κοντινότερος Γείτονας	Συσταδοποίηση
Χρησιμοποιείτε για την πρόβλεψη καθώς επίσης και τη σταθεροποίηση.	Χρησιμοποιείτε συνήθως για την παγίωση των στοιχείων σε μια υψηλού επίπεδου εικόνα και για μια γενική ομαδοποίηση των αρχείων με ομοειδείς συμπεριφορές.
Το διάστημα καθορίζεται από το πρόβλημα για λύση (εποπτευομένη εκμάθηση).	Το διάστημα ορίζεται ως το ν-διάστατο διάστημα προεπιλογής (default n-dimensional space), ή καθορίζεται από το χρήστη, ή είναι ένα προκαθορισμένο διάστημα που οδηγείται από προηγούμενη εμπειρία (ανεπίβλεπτη εκμάθηση).
Γενικά χρησιμοποιεί μόνο μετρήσεις αποστάσεων για να καθορίσει το πόσο κοντά είναι τα δεδομένα.	Μπορεί να χρησιμοποιήσει άλλες μετρήσεις εκτός από την απόσταση για να καθορίσει την απόσταση δύο αρχείων - παραδείγματος χάριν με το να συνδέσει δύο σημεία.

Πίνακας 2.5 *Μερικές από τις διαφορές μεταξύ της τεχνικής εξόρυξης δεδομένων του κοντινότερου γείτονα και της συσταδοποίησης*

Τι είναι ένα ν-διάστατο διάστημα;

Όταν οι άνθρωποι μιλούν για τη συσταδοποίηση ή για την πρόβλεψη κοντινότερου γείτονα θα μιλήσουν συχνά για ένα διάστημα "N" διαστάσεων. Αυτό που εννοούν είναι ότι προκειμένου να καθοριστεί ποιο είναι κοντά και ποιο είναι μακριά είναι χρήσιμο να καθοριστεί ένα διάστημα όπου η απόσταση μπορεί να υπολογιστεί. Γενικά αυτά τα διαστήματα συμπεριφέρονται ακριβώς όπως το τρισδιάστατο διάστημα που γνωρίζουμε, όπου η απόσταση μεταξύ των αντικειμένων καθορίζεται από την ευκλείδεια απόσταση (ακριβώς όπως τον υπολογισμό του μήκους μιας πλευράς σε ένα τρίγωνο).

Αυτά που συμβαίνουν στις τρεις διαστάσεις δουλεύουν επίσης αρκετά καλά και για περισσότερες διαστάσεις. Το οποίο είναι καλό αφού τα περισσότερα πραγματικά παγκόσμια προβλήματα αποτελούνται από πολύ περισσότερες από τρεις διαστάσεις. Στην πραγματικότητα κάθε προάγγελος (ή στήλη βάσεων δεδομένων) που χρησιμοποιείτε μπορεί να θεωρηθεί ως μια νέα διάσταση. Στο παράδειγμα παραπάνω από τους πέντε προάγγελοι (predictors): ηλικία, έσοδα, εισόδημα, μάτια και φύλο μπορούν όλοι να αναλυθούν για να είναι διαστάσεις σε ένα N-διάστατο διάστημα όπου ν, σε αυτήν την περίπτωση είναι 5. Είναι ευκολότερο μερικές φορές να σκεφτεί κάποιος για αυτούς και για άλλους αλγορίθμους εξόρυξης δεδομένων από την άποψη των ν-διάστατων διαστημάτων επειδή επιτρέπει τη χρησιμοποίηση μερικών διαισθήσεων για το πώς λειτουργεί ο αλγόριθμος .

Η μετακίνηση από τρεις διαστάσεις σε πέντε διαστάσεις δεν είναι ένα πάρα πολύ μεγάλο άλμα αλλά υπάρχουν επίσης διαστήματα στα πραγματικά παγκόσμια προβλήματα που είναι πολύ πιο σύνθετα. Στην βιομηχανία πιστωτικών καρτών, οι εκδότες πιστωτικών καρτών, συνήθως έχουν πάνω από χίλιους προάγγελους που θα μπορούσαν να χρησιμοποιηθούν για να δημιουργήσουν ένα ν-διάστατο διάστημα. Για την ανάκτηση κειμένων (π.χ. βρίσκοντας τα χρήσιμα άρθρα στη Wall Street Journal από μια μεγάλη βάση δεδομένων, ή βρίσκοντας τους χρήσιμους ιστοχώρους στο διαδίκτυο) οι προάγγελοι (και ως εκ τούτου οι διαστάσεις) είναι συνήθως λέξεις ή φράσεις που βρίσκονται στα αρχεία εγγράφων. Σε ένα μόνο έτος της Wall Street Journal χρησιμοποιούνται περισσότερες από 50.000 διαφορετικές λέξεις - που μεταφράζεται σε ένα 50.000διάστατο διάστημα στο οποίο οι αποστάσεις μεταξύ των αρχείων πρέπει να υπολογιστούν.

Πώς καθορίζεται το διάστημα για τη συσταδοποίηση και τον κοντινότερο γείτονα;

Για τη συσταδοποίηση το ν-διάστατο διάστημα καθορίζεται συνήθως με την εκχώρηση ενός προάγγελου σε κάθε διάσταση. Για τον αλγόριθμο κοντινότερου γείτονα οι

προάγγελοι χαρτογραφούνται επίσης σε διαστάσεις αλλά έπειτα εκείνες οι διαστάσεις κυριολεκτικά τεντώνονται ή συμπιέζονται σύμφωνα με το πόσο σημαντικός είναι ο συγκεκριμένος προάγγελος στην παραγωγή της πρόβλεψης. Το τέντωμα μιας διάστασης κάνει αποτελεσματικά σημαντικότερη εκείνη την διάσταση (και ως εκ τούτου τον προάγγελο) από τις άλλες στον υπολογισμό της απόστασης.

Για παράδειγμα εάν είστε ορειβάτης βουνών και κάποιος σας είπε ότι ήσαστε 2 μίλια από τον προορισμό σας, η απόσταση είναι η ίδια εάν είναι 1 μίλι βόρεια και 1 μίλι επάνω στο βουνό ή 2 μίλια βόρεια σε επίπεδο έδαφος αλλά σαφώς η προηγούμενη διαδρομή είναι πολύ διαφορετική από ότι η τελευταία. Η απόσταση που διανύετε κατευθείαν προς τα πάνω είναι η σημαντικότερη για να υπολογίσετε πόσο καιρό θα πάρει πραγματικά για να φτάσετε στον προορισμό και θα επιθυμούσατε πιθανώς να θεωρήσετε αυτήν την "διάσταση" σημαντικότερη από άλλες. Στην πραγματικότητα, ως ορειβάτης βουνών, θα μπορούσατε "να σταθμίσετε" τη σημασία της κάθετης διάστασης στον υπολογισμό κάποιας νέας απόστασης με το συλλογισμό ότι κάθε μίλι προς τα πάνω είναι ισοδύναμο με 10 μίλια σε επίπεδο έδαφος.

Εάν χρησιμοποιήσατε αυτήν την εμπειροτεχνική μέθοδο για να σταθμίσετε τη σημασία μιας διάστασης από την άλλη θα ήταν σαφές ότι στη μια περίπτωση ήσαστε πολύ "πιο μακριά" από τον προορισμό σας ("11 μίλια") απ' ότι στο δεύτερο ("2 μίλια").

2.8.2 Ιεραρχική και μη-ιεραρχική συσταδοποίηση

Σύμφωνα με τους Alex Berson, Stephen Smith και Kurt Thearling (1999), υπάρχουν δύο κύριοι τύποι τεχνικών συσταδοποίησης, εκείνων που δημιουργούν μια ιεραρχία των συστάδων και εκείνων που δεν δημιουργούν. Οι τεχνικές ιεραρχικής συσταδοποίησης, δημιουργούν μια ιεραρχία των συστάδων από την πιο μικρή στην πιο μεγάλη. Ο κύριος λόγος για αυτό είναι ότι, όπως έχουμε ήδη πει, η συσταδοποίηση είναι μια ανεπίβλεπτη τεχνική εκμάθησης, και υπό αυτήν τη μορφή, δεν υπάρχει καμία απολύτως σωστή απάντηση. Για αυτόν τον λόγο και ανάλογα με την ιδιαίτερη εφαρμογή της συσταδοποίησης, μικρότερος ή μεγαλύτερος αριθμός συστάδων μπορεί να επιδιωχτεί. Με μια καθορισμένη ιεραρχία των συστάδων είναι δυνατό να επιλεχτεί ο αριθμός των συστάδων που επιδιώκονται. Είναι δυνατό να υπάρξουν τόσες συστάδες όσα είναι τα αρχεία στη βάση δεδομένων. Σε αυτήν την περίπτωση τα αρχεία μέσα στη συστάδα είναι βέλτιστα παρόμοια το ένα με το άλλο (δεδομένου ότι υπάρχει μόνο μία) και βεβαίως διαφορετικά από τις άλλες συστάδες. Αλλά φυσικά μια τέτοια τεχνική συσταδοποίησης χάνει το νόημα υπό την έννοια ότι η ιδέα της συσταδοποίησης είναι να βρεθούν χρήσιμα πρότυπα στη βάση δεδομένων που την

συνοψίζουν και καθιστούν ευκολότερο να τη καταλάβουμε. Οποιοσδήποτε αλγόριθμος συσταδοποίησης που καταλήγει με τόσες συστάδες όσα είναι τα αρχεία δεν βοηθάει το χρήστη να καταλάβει τα στοιχεία καλύτερα. Κατά συνέπεια ένα από τα κύρια σημεία για τη συσταδοποίηση είναι ότι υπάρχουν πολύ λιγότερες συστάδες από τα αρχικά αρχεία. Ακριβώς πόσες συστάδες πρέπει να διαμορφωθούν είναι ένα θέμα ερμηνείας. Το πλεονέκτημα των μεθόδων ιεραρχικής συσταδοποίησης είναι ότι επιτρέπουν στον τελικό χρήστη να επιλέξει είτε πολλές συστάδες είτε μερικές.

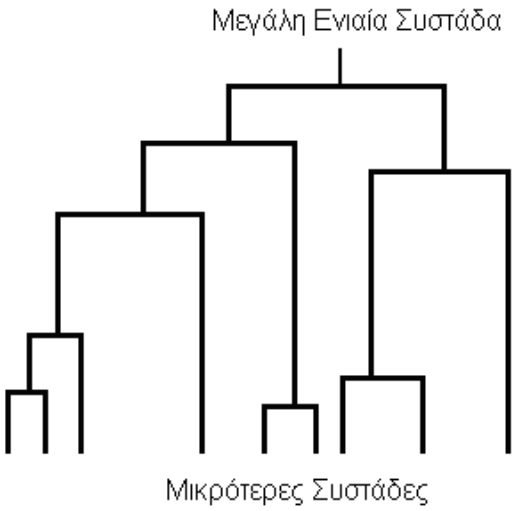
Η ιεραρχία των συστάδων αντιμετωπίζεται συνήθως ως δέντρο όπου οι μικρότερες συστάδες συγχωνεύονται μαζί για να δημιουργήσουν το επόμενο πιο υψηλό επίπεδο συστάδων και εκείνων που σε εκείνη την συγχώνευση επιπέδων δημιουργούν μαζί το επόμενο πιο υψηλό επίπεδο συστάδων. Το σχήμα 2.3 παρακάτω επιδεικνύει πώς διάφορες συστάδες μπορούν να διαμορφώσουν μια ιεραρχία. Όταν δημιουργείται μια τέτοια ιεραρχία συστάδων, ο χρήστης μπορεί να καθορίσει ποιος είναι ο σωστός αριθμός συστάδων που συνοψίζει επαρκώς τα στοιχεία ενώ παρέχει ακόμα τις χρήσιμες πληροφορίες (στο άλλο άκρο μια ενιαία συστάδα που περιέχει όλα τα αρχεία είναι μια μεγάλη περιληπτική παρουσίαση της πληροφορίας αλλά δεν περιέχει αρκετές συγκεκριμένες πληροφορίες για να είναι χρήσιμη).

Αυτή η ιεραρχία των συστάδων δημιουργείται μέσω του αλγορίθμου που χτίζει τις συστάδες. Υπάρχουν δύο κύριοι τύποι αλγορίθμων ιεραρχικής συσταδοποίησης:

- **Συσωρευτικοί** - οι συσωρευτικές τεχνικές συσταδοποίησης αρχίζουν με τόσες συστάδες όσα είναι τα αρχεία όπου κάθε συστάδα περιέχει μόνο ένα αρχείο. Οι συστάδες που είναι πλησιέστερα μεταξύ τους συγχωνεύονται για να διαμορφώσουν την επόμενη μεγαλύτερη συστάδα. Αυτή η συγχώνευση συνεχίζεται έως ότου χτίζεται μια ιεραρχία συστάδων με ακριβώς μια ενιαία συστάδα που περιέχει όλα τα αρχεία στην κορυφή της ιεραρχίας.
- **Διαιρετικοί** - οι διαιρετικές τεχνικές συσταδοποίησης υιοθετούν την αντίθετη μέθοδο από τις συσωρευτικές τεχνικές. Αυτές οι τεχνικές αρχίζουν με όλα τα αρχεία σε μια συστάδα και προσπαθούν έπειτα να χωρίσουν εκείνη την συστάδα σε μικρότερα κομμάτια και έπειτα προσπαθούν να χωρίσουν εκείνα τα μικρότερα κομμάτια.

Από τις δύο, οι συσωρευτικές τεχνικές είναι που χρησιμοποιούνται συχνότερα για τη συσταδοποίηση και έχουν περισσότερους αλγόριθμους να αναπτύσσονται για αυτές. Οι μη-ιεραρχικές τεχνικές μπορούν να δημιουργηθούν γρηγορότερα από την βάση δεδομένων αλλά απαιτούν από το χρήστη να λάβει κάποια απόφαση για τον αριθμό των επιθυμητών συστάδων ή την ελάχιστη "εγγύτητα" που απαιτείται για δύο αρχεία για να είναι μέσα στην ίδια

συστάδα. Αυτές οι μη-ιεραρχικές τεχνικές συχνά τρέχουν πολλές φορές, αρχίζοντας με κάποια αυθαίρετη ή ακόμα και τυχαία συσταδοποίηση και έπειτα επαναληπτικά βελτιώνοντας τη συσταδοποίηση με το να μεταθέσουν μερικά αρχεία τριγύρω. Αυτές οι τεχνικές μερικές φορές δημιουργούν τις συστάδες που δημιουργούνται με μόνο ένα πέρασμα μέσω της βάσης δεδομένων προσθέτοντας τα αρχεία στις υπάρχουσες συστάδες όταν υπάρχουν και δημιουργώντας τις νέες συστάδες όταν καμία υπάρχουσα συστάδα δεν είναι καλός υποψήφιος για το δεδομένο αρχείο. Επειδή ο καθορισμός των ποιων συστάδων θα διαμορφωθούν, μπορεί να εξαρτηθεί από αυτές τις αρχικές επιλογές των οποίων οι αρχικές συστάδες πρέπει να επιλεχτούν ή ακόμα και σε πόσες συστάδες αυτές οι τεχνικές μπορούν να είναι λιγότερο επαναλαμβανόμενες από τις ιεραρχικές τεχνικές και μπορούν μερικές φορές να δημιουργήσουν είτε πάρα πολλές είτε πολύ λίγες συστάδες, επειδή ο αριθμός συστάδων προκαθορίζεται από το χρήστη και δεν καθορίζεται απλώς από τα έμφυτα σχέδια στη βάση δεδομένων.



Σχήμα 2.3 Διάγραμμα που παρουσιάζει μια ιεραρχία των συστάδων. Οι συστάδες στο χαμηλότερο επίπεδο συγχωνεύονται μαζί για να διαμορφώσουν τις μεγαλύτερες συστάδες στο επόμενο επίπεδο της ιεραρχίας

Μη-ιεραρχική συσταδοποίηση

Υπάρχουν δύο κύριες τεχνικές μη-ιεραρχικής συσταδοποίησης. Και οι δύο είναι πολύ γρήγορες στον υπολογισμό της βάσης δεδομένων αλλά έχουν μερικά μειονεκτήματα. Οι πρώτες είναι οι μέθοδοι ενός περάσματος (single pass methods). Αντλούν το όνομά τους από το γεγονός ότι η βάση δεδομένων πρέπει να περαστεί μια φορά προκειμένου να δημιουργηθούν οι συστάδες (δηλ. κάθε αρχείο διαβάζεται μόνο μία φορά από τη βάση δεδομένων). Η άλλη κατηγορία τεχνικών ονομάζεται μέθοδοι ανακατανομής (reallocation methods). Παίρνουν το όνομά τους από τη μετακίνηση ή την "ανακατανομή" των αρχείων από μια συστάδα σε άλλη προκειμένου να δημιουργηθούν καλύτερες συστάδες. Οι τεχνικές ανακατανομής χρησιμοποιούν πολλαπλά περάσματα μέσω της βάσης δεδομένων αλλά είναι σχετικά γρήγορες σε σύγκριση με τις ιεραρχικές τεχνικές.

Μερικές τεχνικές επιτρέπουν στο χρήστη να ζητήσει τον αριθμό συστάδων που θα επιθυμούσε να εξαχθεί από τα στοιχεία. Το να προκαθορίσουμε τον αριθμό των συστάδων παρά να οδηγηθεί από τα στοιχεία μπορεί να φαίνεται κακή ιδέα δεδομένου ότι μπορεί υπάρξει κάποια πολύ ευδιάκριτη και αισθητή συγκέντρωση των στοιχείων σε ορισμένες συστάδες τις οποίες ο χρήστης να μην γνωρίζει.

Παραδείγματος χάριν ο χρήστης μπορεί να θέλει να δει τα στοιχεία να χωρίζονται σε 10 συστάδες αλλά τα ίδια τα στοιχεία να χωρίζονται πολύ καθαρά σε 13 συστάδες. Αυτές οι μη-ιεραρχικές τεχνικές θα δοκιμάσουν να στριμώξουν αυτές τις πρόσθετες τρεις συστάδες στα υπάρχουσες 10 παρά τη δημιουργία 13 που θα χώριζαν καλύτερα τα δεδομένα. Η εξοικονόμηση χάριτος για αυτές τις μεθόδους, εντούτοις, είναι ότι, όπως έχουμε δει, δεν υπάρχει μόνο μια σωστή απάντηση για το πώς να γίνει η συσταδοποίηση, έτσι είναι σπάνιο με αυθαίρετο προκαθορισμό του αριθμού των συστάδων να καταλήγατε σε λανθασμένη απάντηση. Ένα από τα πλεονεκτήματα αυτών των τεχνικών είναι ότι συχνά ο χρήστης έχει κάποιο προκαθορισμένο επίπεδο περιληπτικής παρουσίασης της πληροφορίας που τον ενδιαφέρει (π.χ. "25 συστάδες είναι πολύ μπερδεμένες, αλλά 10 θα βοηθήσουν να μου δώσουν μια διορατικότητα για τα στοιχεία μου"). Το γεγονός ότι μεγαλύτεροι ή λιγότεροι αριθμοί συστάδων θα ταίριαζαν καλύτερα με τα στοιχεία είναι πραγματικά δευτερεύουσας σημασίας.

Ιεραρχική συσταδοποίηση

Η ιεραρχική συσταδοποίηση έχει το πλεονέκτημα σε σχέση με τις μη-ιεραρχικές τεχνικές, δεδομένου ότι οι συστάδες καθορίζονται απλώς από τα στοιχεία (όχι από τους

χρήστες που προκαθορίζουν τον αριθμό συστάδων) και ότι ο αριθμός συστάδων μπορεί να αυξηθεί ή να μειωθεί με μια απλή κίνηση πάνω-κάτω στην ιεραρχία.

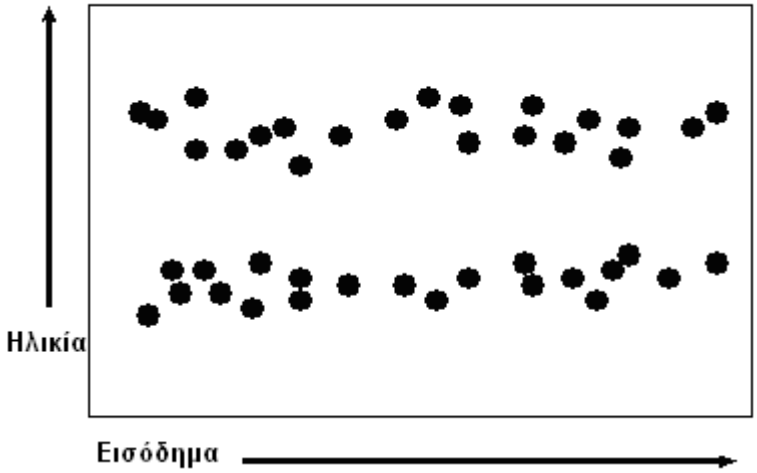
Η ιεραρχία δημιουργείται με την έναρξη είτε στην κορυφή (μια συστάδα που περιλαμβάνει όλα τα αρχεία) και την υποδιαίρεση (διαχωριστική συγκέντρωση) είτε με την έναρξη στο κατώτατο σημείο με τόσες συστάδες όσα είναι τα αρχεία και συγχώνευση (συσωρευτική συγκέντρωση). Συνήθως η συγχώνευση και η υποδιαίρεση γίνονται δύο συστάδες τη φορά.

Η κύρια διάκριση μεταξύ των τεχνικών είναι η δυνατότητά τους να ευνοήσουν μακριές, "απεριποίητες" συστάδες που είναι συνδεμένες αρχείο-αρχείο, ή για να ευνοήσουν την ανίχνευση μιας κλασσικότερης, συμπαγούς ή σφαιρικής συστάδας. Μπορεί να φαίνεται παράξενο να θέλει κάποιος να διαμορφώσει αυτή την μακριά συστάδα, αλλά σε μερικές περιπτώσεις είναι οι "χτύποι" (patters) που ο χρήστης επιθυμεί να έχουν ανιχνευτεί στη βάση δεδομένων. Τότε είναι που το ελλοχεύον διάστημα φαίνεται αρκετά διαφορετικό από τις σφαιρικές συστάδες και οι συστάδες που πρέπει να διαμορφωθούν δεν βασίζονται στην απόσταση από το κέντρο της συστάδας αλλά αντ' αυτού βασίζονται στα αρχεία "που συνδέονται". Εξετάστε το παράδειγμα που παρουσιάζεται στο σχήμα 2.4 ή στο σχήμα 2.5. Σε αυτές τις περιπτώσεις υπάρχουν δύο συστάδες που δεν είναι πολύ σφαιρικές στη μορφή αλλά θα μπορούσαν να ανιχνευθούν από την τεχνική απλού συνδέσμου (single link technique).

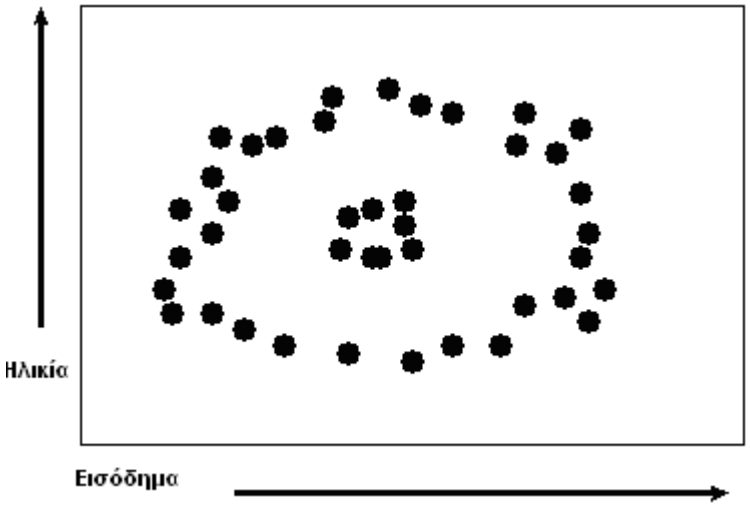
Κατά την εξέταση του σχεδιαγράμματος των στοιχείων στο σχήμα 2.4 φαίνεται να υπάρχουν δύο σχετικά επίπεδες συστάδες που τρέχουν παράλληλα κάθε μια κατά μήκος του εισοδηματικού άξονα. Ούτε η πλήρης σύνδεση ούτε η μέθοδος Ward, εντούτοις, δεν θα επέστρεφε αυτές τις δύο συστάδες στο χρήστη. Αυτές οι τεχνικές στηρίζονται στη δημιουργία ενός "κέντρου" για κάθε συστάδα και την επιλογή αυτών των κέντρων έτσι ώστε κατά μέσο όρο η απόσταση κάθε αρχείου από αυτό το κέντρο να ελαχιστοποιείται. Τα σημεία που είναι πολύ μακριά από αυτά τα κέντρα θα περιέρχονταν απαραιτήτως σε μια διαφορετική συστάδα.

Αυτό που καθιστά αυτές τις συστάδες "ορατές" σε αυτό το απλό δυσδιάστατο διάστημα είναι το γεγονός ότι κάθε σημείο σε μια συστάδα συνδέεται στενά με κάποιο άλλο σημείο. Για τις δύο συστάδες βλέπουμε ότι η μέγιστη απόσταση μεταξύ των κοντινότερων δύο σημείων μέσα σε μια συστάδα είναι μικρότερη από την ελάχιστη απόσταση των κοντινότερων δύο σημείων στις διαφορετικές συστάδες. Δηλαδή ότι για οποιοδήποτε σημείο σε αυτό το διάστημα, το κοντινότερο σημείο σε αυτό πρόκειται πάντα να είναι ένα άλλο

σημείο στην ίδια συστάδα. Τώρα το κέντρο βάρους μιας συστάδας θα μπορούσε να είναι αρκετά μακριά από ένα δεδομένο σημείο αλλά ότι κάθε σημείο συνδέεται με κάθε άλλο σημείο από μια σειρά μικρών αποστάσεων. (Alex Berson, Stephen Smith, Kurt Thearling , 1999)



Σχήμα 2.4 *ένα παράδειγμα των επιμηκυμένων συστάδων που δεν θα ανακτώνταν με τη μέθοδο της πλήρης σύνδεσης ή τη μέθοδο Ward αλλά με τη μέθοδο της μονής σύνδεσης.*



Σχήμα 2.5 *Ένα παράδειγμα των τοποθετημένων συστάδων που δεν θα ανακτώνταν με τις μεθόδους πλήρης σύνδεσης ή της μεθόδου Ward αλλά με τη μέθοδο μονής σύνδεσης.*

2.9 Ανάλυση συνδέσεων (link analysis)

Το διαδίκτυο (World Wide Web) είναι μια πλούσια πηγή πληροφοριών και συνεχίζει να επεκτείνετε σε μέγεθος και πολυπλοκότητα. Η ανάκτηση της απαιτούμενης από το χρήστη ιστοσελίδας στον Ιστό, αποδοτικά και αποτελεσματικά, είναι μια πρόκληση. Όταν ένας χρήστης θέλει να ψάξει για κάποιες σελίδες, προτιμά αυτές οι σελίδες να είναι προσιτές. Γίνεται πολύ δύσκολο για τους χρήστες να βρουν, να εξάγουν, να φιλτράρουν ή να αξιολογήσουν τις σχετικές πληροφορίες λόγο του μεγάλου αριθμού πληροφοριών. Αυτό το ζήτημα δημιουργεί την ανάγκη για κάποια τεχνική που θα μπορεί να λύσει αυτές τις προκλήσεις. Η εξόρυξη γνώσης στον Παγκόσμιο Ιστό μπορεί να εκτελεστεί εύκολα με τη βοήθεια της βάση δεδομένων (Database, DB), της ανάκτησης πληροφοριών (Information retrieval , IR), της επεξεργασίας φυσικής γλώσσας (Natural Language Processing, NLP), και της μηχανικής μάθησης (Machine Learning) κλπ. (Tamanna Bhatia, 2011)
Οι προκλήσεις στη εξόρυξη γνώσης στον Παγκόσμιο Ιστό είναι:

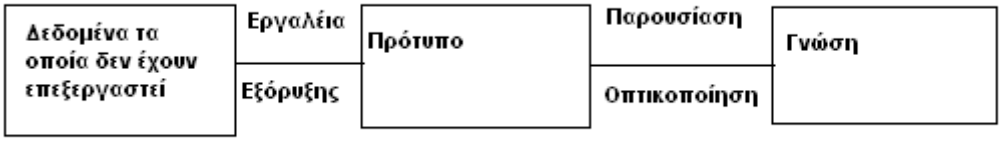
1. Ο Ιστός είναι τεράστιος.
2. Οι Ιστοσελίδες είναι ημι-δομημένες.
3. Οι πληροφορίες Ιστού παρουσιάζουν ποικιλομορφία στη σημασία.
4. Ο βαθμός ποιότητας των πληροφοριών που εξάγονται.
5. Το συμπέρασμα της γνώσης από τις πληροφορίες που εξάγονται.

2.9.1. Εξόρυξη γνώσης στον Παγκόσμιο Ιστό

Η εξόρυξη γνώσης στον Παγκόσμιο Ιστό είναι η τεχνική εξόρυξης δεδομένων που αυτόματα ανακαλύπτει ή εξάγει τις πληροφορίες από τα έγγραφα Ιστού. Συγκεκριμένα είναι η εξαγωγή των ενδιαφέρων και των ενδεχομένως χρήσιμων σχεδίων και πληροφοριών από αντικείμενα ή δραστηριότητες σχετικές με το διαδίκτυο.

Διαδικασία εξόρυξης γνώσης στον παγκόσμιο ιστό

Η πλήρης διαδικασία της εξόρυξης γνώσης από τα στοιχεία του παγκόσμιου Ιστού φαίνεται στο σχήμα 2.6:



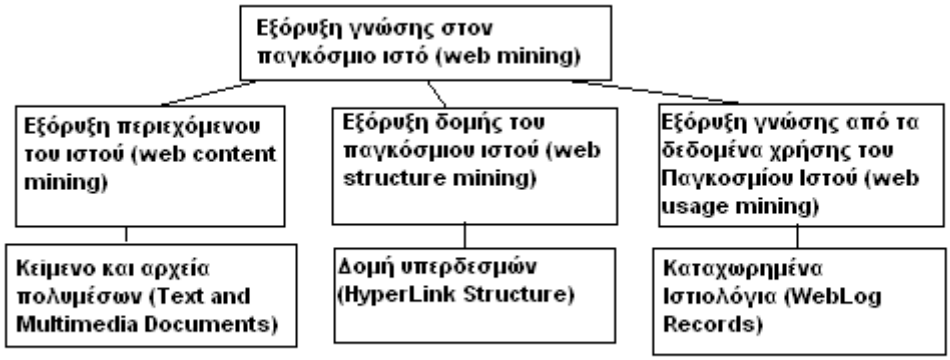
Σχήμα 2.6 : Διαδικασία εξόρυξης γνώσης στον παγκόσμιο ιστό

Τα διάφορα βήματα εξηγούνται ως εξής.

1. Εύρεση πόρων: Είναι η εργασία ανεύρεσης αρχείων στο διαδίκτυο
2. Επιλογή πληροφοριών και προεπεξεργασία: Αυτόματη επιλογή και προ-επεξεργασία συγκεκριμένων πληροφοριών από τις πληροφορίες που ανακτήθηκαν από τους πόρους του διαδικτύου.
3. Γενίκευση: Αυτόματα ανακαλύπτει τα γενικά πρότυπα ενός ιστοχώρου καθώς επίσης και πολλαπλών ιστοχώρων.
4. Ανάλυση: Επικύρωση και ερμηνεία των εξαγμένων προτύπων.

Κατηγορίες εξόρυξης γνώσης στον παγκόσμιο ιστό

Η έρευνα για την εξόρυξη γνώσης στον παγκόσμιο ιστό συμπίπτει ουσιαστικά με άλλες περιοχές, συμπεριλαμβανομένης της εξόρυξης δεδομένων, της εξόρυξης κειμένων, της ανάκτησης πληροφοριών, και της ανάκτησης Ιστού. Η ταξινόμηση είναι βασισμένη σε δύο πτυχές: τον σκοπό και τις πηγές των στοιχείων. Η έρευνα για την ανάκτηση εστιάζει στην ανάκτηση σχετικών, υπαρκτών στοιχείων ή έγγραφων από μια μεγάλη βάση δεδομένων ή μια αποθήκη εγγράφων, ενώ η ερευνά για την εξόρυξη εστιάζει στην ανακάλυψη νέων πληροφοριών ή γνώσεων στα στοιχεία. Βάσει αυτού, η εξόρυξη γνώσης στον παγκόσμιο ιστό μπορεί να ταξινομηθεί στην εξόρυξη δομής του παγκόσμιου ιστού (web structure mining), εξόρυξη περιεχόμενου του παγκόσμιου ιστού (web content mining), και εξόρυξη γνώσης από τα δεδομένα χρήσης του Παγκοσμίου Ιστού (web usage mining) όπως φαίνεται στο σχήμα 2.7.



Σχήμα 2.7: Κατηγορίες εξόρυξης γνώσης στον παγκόσμιο ιστό

Α. Εξόρυξη περιεχόμενου του παγκόσμιου ιστού (web content mining)

Η εξόρυξη περιεχόμενου του παγκόσμιου ιστού είναι η διαδικασία εξαγωγής χρήσιμων πληροφοριών από το περιεχόμενο των εγγράφων του παγκόσμιου ιστού. Τα στοιχεία αντιστοιχούν στη συλλογή των πληροφοριών μιας ιστοσελίδας που σχεδιάστηκαν για να μεταβιβαστούν στους χρήστες. Η εξόρυξη περιεχόμενου του παγκόσμιου ιστού συσχετίζεται αλλά διαφέρει από την εξόρυξη δεδομένων και τη εξόρυξη κειμένων. Συσχετίζεται με τη εξόρυξη κειμένων επειδή ένα μεγάλο μέρος του περιεχομένου του παγκόσμιου Ιστού είναι κείμενα. Διαφέρει από την εξόρυξη δεδομένων επειδή τα στοιχεία του ιστού είναι κυρίως ημι-δομημένα ή/και μη-δομημένα. Η εξόρυξη περιεχόμενου του παγκόσμιου ιστού διαφέρει επίσης από την εξόρυξη κειμένων λόγω της ημι-δομημένης φύσης του ιστού, ενώ η εξόρυξη κειμένων εστιάζει στα μη-δομημένα κείμενα. Οι τεχνολογίες που συνήθως χρησιμοποιούνται στη εξόρυξη περιεχομένου του παγκόσμιου ιστού είναι η Natural language processing(NLP) (Επεξεργασία φυσικής γλώσσας) και Information retrieval (IR) (ανάκτηση πληροφοριών).

Β. Εξόρυξη δομών του παγκόσμιου ιστού (web structure mining)

Η εξόρυξη δομών του παγκόσμιου ιστού είναι η διαδικασία από την οποία ανακαλύπτουμε το πρότυπο της δομής των συνδέσεων των ιστοσελίδων . Ο στόχος της εξόρυξης δομών του παγκόσμιου ιστού είναι να παράγει τη δομημένη περίληψη για τον ιστοχώρο και την ιστοσελίδα. Προσπαθεί να ανακαλύψει τη δομή των συνδέσεων των υπερσυνδέσεων σε επίπεδο εγγράφων. Το άλλο είδος της εξόρυξης δομών του παγκόσμιου ιστού εξάγει την δομή εγγράφων. Χρησιμοποιεί τη δεντρόμορφη δομή για να αναλύσει και

να περιγράψει το HTML (Hyper Text Markup Language) (Γλώσσα Σήμανσης Υπερκειμένου) ή XML (Extensible Markup Language) (Επεκτάσιμη γλώσσα σήμανσης).

Γ. Εξόρυξη γνώσης από τα δεδομένα χρήσης του Παγκοσμίου Ιστού (web usage mining)

Η εξόρυξη γνώσης από τα δεδομένα χρήσης του Παγκοσμίου Ιστού είναι η διαδικασία από την οποία προσδιορίζουμε τα πρότυπα φυλλομέτρησης με την ανάλυση της συμπεριφοράς της πλοήγησης του χρήστη. Εστιάζει στις τεχνικές που μπορούν να χρησιμοποιηθούν για να προβλέψουν τη συμπεριφορά των χρηστών ενώ ο χρήστης αλληλεπιδρά με τον ιστό. Χρησιμοποιεί δευτεροβάθμια στοιχεία όσον αφορά τον ιστό. Αυτή η δραστηριότητα περιλαμβάνει την αυτόματη ανακάλυψη των προτύπων πρόσβασης των χρηστών από έναν ή περισσότερους διακομιστές (web server). Μέσω αυτής της τεχνικής εξόρυξης μπορούμε να εξακριβώσουμε τι αναζητούν οι χρήστες στο διαδίκτυο. Αποτελείται από τρεις φάσεις, την προ-επεξεργασία, την ανακάλυψη προτύπων, και την ανάλυση προτύπων. Οι διακομιστές,, οι proxies και οι εφαρμογές πελατών μπορούν αρκετά εύκολα να συλλάβουν τα στοιχεία για τη χρήση του παγκόσμιου ιστού.

2.9.2. Αλγόριθμοι ανάλυσης συνδέσεων

Η τεχνική εξόρυξης γνώσης στον παγκόσμιο ιστό παρέχει τις πρόσθετες πληροφορίες μέσω υπερσυνδέσεων όπου διαφορετικά έγγραφα συνδέονται. Μπορούμε να δούμε τον Ιστό ως μια κατευθυνόμενη γραφική παράσταση, οι κόμβοι της οποίας είναι τα έγγραφα ή οι σελίδες και οι άκρες είναι οι υπερσυνδέσεις μεταξύ τους. Αυτή η κατευθυνόμενη δομή γραφικών παραστάσεων είναι γνωστή ως γραφική παράσταση του ιστού. Υπάρχει πλήθος αλγορίθμων βασισμένων στην ανάλυση συνδέσεων. Οι τρεις σημαντικότεροι αλγόριθμοι είναι: Page Rank, Weighted Page Rank και Weighted Page Content Rank.(Tamanna Bhatia, 2011)

A. Page Rank

Το Page Rank είναι μια αριθμητική αξία που αντιπροσωπεύει το πόσο σημαντική είναι μια σελίδα στον Ιστό. Είναι η μέθοδος που χρησιμοποιεί το Google για τη μέτρηση της σημαντικότητας μιας σελίδας. Όταν όλοι οι άλλοι παράγοντες όπως ο τίτλος (Titletag) και οι λέξεις-κλειδιά (keywords) λαμβάνονται υπόψη, το Google χρησιμοποιεί το Page Rank για να ρυθμίσει τα αποτελέσματα έτσι ώστε οι περισσότερο "σημαντικές" σελίδες να μετακινούνται επάνω στη σελίδα αποτελεσμάτων στη οθόνη αναζήτησης ενός χρήστη. Το Google αντιλαμβάνεται ότι όταν μια σελίδα συνδέεται με μια άλλη, αυτή προσθέτει μια "ψήφο" στην

άλλη σελίδα. Υπολογίζει τη σημασία μιας σελίδας από τις ψήφους που έχει δεχτεί και το πόσο σημαντική είναι η κάθε ψήφος λαμβάνεται υπόψη όταν υπολογίζεται το PageRank της σελίδας. Δεν είναι ο μόνος παράγοντας που χρησιμοποιεί το Google άλλα είναι ένας από τους σημαντικότερους παράγοντες που καθορίζουν τη ταξινόμηση μιας σελίδας στα αποτελέσματα αναζήτησης. Η σειρά ταξινόμησης στο Google λειτουργεί κάπως έτσι:

- ✓ Βρίσκει όλες τις σελίδες που ταιριάζουν με τις λέξεις-κλειδιά της αναζήτησης.
- ✓ Ρυθμίζει τα αποτελέσματα από τα αποτελέσματα του PageRank.

Ο αλγόριθμος του PageRank είναι ο εξής:

Το PageRank λαμβάνει υπόψη τα backlinks και διαδίδει την ταξινόμηση μέσω των συνδέσεων. Μια σελίδα είναι σε υψηλότερη βαθμίδα, εάν το ποσό των τάξεων των backlinks είναι υψηλό. Σχήμα 2.8 : παρουσιάζουν ένα παράδειγμα των backlinks όπου η σελίδα Α είναι ένα backlink της σελίδας Β και της σελίδας Γ ενώ η σελίδα Β και η σελίδα Γ είναι backlinks της σελίδας Δ.

Ο αλγόριθμος του Page Rank δίνεται στην ακόλουθη εξίσωση:

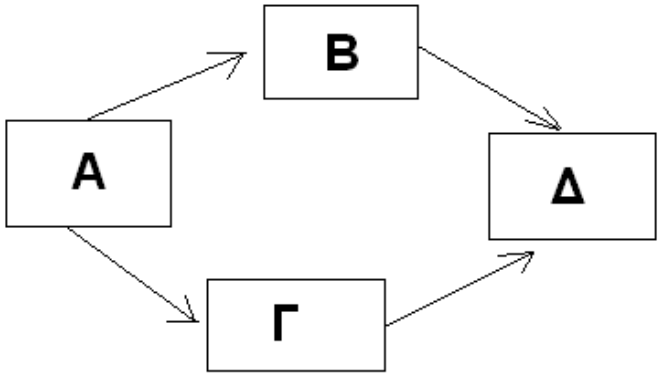
PR(P)=(1-d)+d(PR(T1)/C(T1)+…..PR(Tn)/C(Tn)) ....

Όπου, PR (P)= PageRank της σελίδας P

PR (Ti) = PageRank της σελίδας Ti που οδηγεί με link στη σελίδα

C (Ti) = αριθμός των εξερχόμενων συνδέσεων στη σελίδα T

D = Ο Συντελεστής Απόσβεσης που μπορεί να οριστεί από 0 ως1.



Σχήμα 2.8: Παράδειγμα των backlinks

Β. Σταθμισμένο Page Rank (Weighted Page Rank)

Το σταθμισμένο Page Rank ορίζει την μεγάλης αξίας κατάταξη στις σημαντικότερες σελίδες αντί της διαίρεσης της αξίας της κατάταξης μιας σελίδας ομοιόμορφα μεταξύ των outlink σελίδων της. Η σημαντικότητα ορίζεται από την άποψη του "βάρους" των εισερχόμενων και εξερχόμενων συνδέσεων που γίνονται και αντίστοιχα υπολογίζεται βάσει του αριθμού των εισερχόμενων συνδέσεων με τη σελίδα v και τον αριθμό των εισερχόμενων συνδέσεων με όλες τις σελίδες αναφοράς της σελίδας μ. In είναι ο αριθμός των εισερχόμενων συνδέσεων της σελίδας v, Ip είναι αριθμός των εισερχόμενων συνδέσεων της σελίδας p, R(m) είναι η λίστα των σελίδων αναφοράς της σελίδας m. Υπολογίζεται βάσει του αριθμού εξερχόμενων συνδέσεων της σελίδας n και του αριθμού εξερχόμενων συνδέσεων όλων των σελίδων αναφοράς της σελίδας m. On είναι ο αριθμός εξερχόμενων συνδέσεων της σελίδας n, Op είναι αριθμός εξερχόμενων συνδέσεων της σελίδας p. Έπειτα το σταθμισμένο Page Rank δίνεται με τον ακόλουθο τύπο:

WPR(n)=(1-d)+d

Γ. Σταθμισμένο περιεχόμενο Page Rank (Weighted Page Content Rank)

Ο αλγόριθμος σταθμισμένου περιεχόμενου PageRank (Weighted Page Content Rank Algorithm WPCR) είναι ένας αλγόριθμος κατάταξης σελίδων που χρησιμοποιείται για να δώσει μια ταξινόμηση στις ιστοσελίδες που εμφανίζονται σε μια μηχανή αναζήτησης, σαν απάντηση στην ερώτηση ενός χρήστη. Το WPCR είναι μια αριθμητική αξία που βασίζεται σε ποια ιστοσελίδα δίνεται η διαταγή. Αυτός ο αλγόριθμος υιοθετεί τη εξόρυξη δομών του παγκόσμιου ιστού όπως επίσης και τις τεχνικές εξόρυξης περιεχομένου στον παγκόσμιο ιστό. Η εξόρυξη δομής του παγκόσμιου ιστού χρησιμοποιείται για να υπολογίσει την σημαντικότητα μιας σελίδας και η εξόρυξη περιεχομένου του παγκόσμιου ιστού χρησιμοποιείται για να βρεθεί πόσο σχετική είναι μία σελίδα. Σημαντικότητα είναι η "δημοτικότητα" της σελίδας, δηλαδή πόσες σελίδες δείχνουν ή αναφέρονται σε αυτήν τη συγκεκριμένη σελίδα. Μπορεί να υπολογιστεί βάσει του αριθμού των inlinks και outlinks της σελίδας. Σχετικότητα είναι το ταίριασμα της σελίδας με τα προκαθορισμένα αιτήματα. Αν μια σελίδα ταιριάζει με τα αιτήματα τότε γίνεται πιο σχετική.

Αλγόριθμος: Υπολογιστής WPCR

Input: Σελίδα P, «βάρη» των Inlink και Outlink όλων των backlinks της P, αίτημα Q, d (damping factor) Συντελεστής Απόσβεσης.

Output: Σκορ κατάταξης

Βήμα 1: Υπολογισμός σχετικότητας:

α) Βρείτε όλες τις σημαντικές σειρές λέξης του q (N)

β) Βρείτε εάν οι σειρές ν εμφανίζονται στο π ή όχι;

Z = ποσό των συχνοτήτων όλων των σειρών ν.

γ) C = σύνολο των μέγιστων πιθανών σειρών που εμφανίζονται στο ΡΠ.

δ) X = Σύνολο των συχνοτήτων των σειρών στο S.

ε) Βάρος περιεχομένου (CW) = X/Z

ζ) C = αριθμός των όρων αιτήματος στο P

η) D = αριθμός όλων των όρων αιτήματος του Q αγνοώντας τις λέξεις στάσεων.

θ) Βάρος πιθανότητας (PW) = C/D

Βήμα 2: Υπολογισμός κατάταξης:

α) Βρείτε όλα τα backlinks του P (καθορισμένο B).

β) PR(P)=(1-d)+d[

γ) Απόδοση PR (P) δηλ. το σκορ κατάταξης

2.9.3 Σύγκριση των αλγορίθμων

Ο πίνακας 2.6 παρουσιάζει τη διαφορά μεταξύ των παραπάνω τριών αλγορίθμων:

Πίνακας 2.6: Σύγκριση του PageRank, σταθμισμένου PageRank και σταθμισμένου περιεχόμενου PageRank.(Tamanna Bhatia, 2011)

Περιεχόμενα	Σύγκριση		
	Page Rank	Σταθμισμένο PageRank	Σταθμισμένο περιεχόμενοPageRank
Τεχνική εξόρυξης που χρησιμοποιείται	WSM	WSM	WSM and WCM
Πολυπλοκότητα	O(logn)	<O(logn)	<O(logn)
Εργασιακή διαδικασία	Υπολογίζει τα αποτελέσματα σε indextime. Τα αποτελέσματα ταξινομούνται με τη σημαντικότητα των σελίδων.	Ορίζει μεγάλη αξία στις σημαντικότερες σελίδες αντί να ρίχνει την αξία κατάταξης ομοιόμορφα μεταξύ των outlink σελίδων της.	Δίνει μια ταξινόμηση στις ιστοσελίδες που εμφανίζονται σε μια μηχανή αναζήτησης, σαν απάντηση σε μια ερώτηση ενός χρήστη.
Παράμετροι input/output	Backlinks	Backlinks και forward links	Backlinks, forward links και περιεχόμενο
Πλεονεκτήματα	Παρέχει σημαντικές πληροφορίες για τη δεδομένη ερώτηση με το να ρίξει την αξία κατάταξης εξίσου μεταξύ των	Παρέχει σημαντικές πληροφορίες για τη δεδομένη ερώτηση και ορίζοντας τη σημασία από την άποψη των τιμών	Παρέχει σημαντικές πληροφορίες και σχετικότητα για μια δεδομένη ερώτηση με τη χρήση της εξόρυξης δομής του

	outlink σελίδων της	βάρους σε εισερχόμενες και εξερχόμενες συνδέσεις	παγκόσμιου Ιστού και της εξόρυξης περιεχομένου του παγκόσμιου ιστού
Μηχανή αναζήτησης	Google	Google	Research Model
Περιορισμοί	(1) Το PageRank διανέμεται εξίσου στις εξερχόμενες συνδέσεις (2) Βασίζεται καθαρά στον αριθμό inlinks και outlinks.	(1) Ακόμα και αν μερικές σελίδες μπορεί να είναι άσχετες με μια δεδομένη ερώτηση, λαμβάνουν την υψηλότερη βαθμίδα (2) Υπάρχει ένας λιγότερος προσδιορισμός της σχετικότητας των σελίδων λαμβάνοντας υπόψη την ερώτηση	Κανένας περιορισμός σε σχέση με το PageRank και το σταθμισμένο PageRank

2.10 Γενετικοί αλγόριθμοι (genetic algorithms)

Σύμφωνα με τον David L. Olson & τον Dursun Delen (2008), οι γενετικοί αλγόριθμοι είναι μαθηματικές διαδικασίες που χρησιμοποιούν τη διαδικασία της γενετικής κληρονομιάς. Έχουν εφαρμοστεί σε μια ευρεία ποικιλία προβλημάτων ανάλυσης. Η εξόρυξη δεδομένων μπορεί να συνδυάσει την ανθρώπινη κατανόηση με την αυτόματη ανάλυση των στοιχείων για να ανιχνεύσει τα σχέδια ή τις βασικές σχέσεις. Λαμβάνοντας υπόψη μια μεγάλη βάση δεδομένων που καθορίζεται πέρα από διάφορες μεταβλητές, ο στόχος είναι να βρεθούν αποτελεσματικά τα πιο ενδιαφέροντα σχέδια στη βάση δεδομένων.

Οι γενετικοί αλγόριθμοι έχουν εφαρμοστεί για να προσδιορίσουν ενδιαφέρων σχέδια σε μερικές εφαρμογές. Συνήθως χρησιμοποιούνται στην εξόρυξη δεδομένων για να βελτιώσουν την απόδοση άλλων αλγορίθμων, όπως για παράδειγμα των αλγόριθμων δέντρων απόφασης. Οι γενετικοί αλγόριθμοι απαιτούν ορισμένη δομή δεδομένων. Λειτουργούν σε έναν πληθυσμό που τα χαρακτηριστικά του εκφράζονται με κατηγορική μορφή. Η αναλογία με τη γενετική είναι ότι ο πληθυσμός (γονίδια) αποτελείται από τα χαρακτηριστικά (αλληλόμορφα γονίδια). Ένας τρόπος να εφαρμοστούν οι γενετικοί αλγόριθμοι είναι να εφαρμοστούν οι χειριστές (αναπαραγωγή, διασταύρωση, επιλογή) με το χαρακτηριστικό γνώρισμα της μεταλλαγής για να ενισχύσει την παραγωγή των ενδεχομένως καλύτερων συνδυασμών. Η γενετική διαδικασία αλγορίθμου είναι η εξής:

1. Τυχαία επιλογή γονιών.
2. Αναπαραγωγή μέσω της διασταύρωσης. Η αναπαραγωγή είναι η επιλογή του χειριστή για το ποιες μεμονωμένες οντότητες θα επιζήσουν. Με άλλα λόγια, κάποιο χαρακτηριστικό λειτουργίας ή επιλογής απαιτείται για να καθορίσει την επιβίωση.
Η διασταύρωση αφορά τις αλλαγές στις μελλοντικές γενεές των οντοτήτων.
3. Επιλογή των επιζώντων για την επόμενη γενεά μέσω μιας λειτουργίας ικανότητας.
4. Η μεταλλαγή είναι η λειτουργία από την οποία τυχαία επιλεγμένες ιδιότητες από τυχαία επιλεγμένες οντότητες σε επόμενες διαδικασίες αλλάζουν.
5. Επαναλάβετε έως ότου είτε επιτυγχάνεται ένα δεδομένο επίπεδο ικανότητας, είτε ο προκαθορισμένος αριθμός επαναλήψεων επιτυγχάνεται.

Οι παράμετροι του γενετικού αλγορίθμου περιλαμβάνουν το μέγεθος πληθυσμού, ποσοστό διασταυρώσεων (η πιθανότητα που τα άτομα θα διασταυρωθούν), και το ποσοστό μεταλλαγής (η πιθανότητα που μια ορισμένη οντότητα μεταλλάσετε).

Πλεονεκτήματα γενετικού αλγορίθμου: Οι γενετικοί αλγόριθμοι είναι πολύ εύκολο να αναπτυχθούν, το οποίο τους καθιστά ιδιαίτερα ελκυστικούς. Ο αλγόριθμος είναι παράλληλος, σημαίνοντας ότι μπορεί να χρησιμοποιηθεί για μεγάλους πληθυσμούς αποτελεσματικά. Ο αλγόριθμος είναι επίσης αποδοτικός εάν αρχίζει με μια φτωχή αρχική λύση, τότε μπορεί γρήγορα να προχωρήσει στις καλές λύσεις. Η χρήση της μεταλλαγής καθιστά τη μέθοδο ικανή να αναγνωρίσει τα σφαιρικά βέλτιστα ακόμη και στις μη γραμμικές περιοχές του προβλήματος. Η μέθοδος δεν απαιτεί γνώση για τη διανομή των στοιχείων.

Μειονεκτήματα γενετικού αλγορίθμου: Οι γενετικοί αλγόριθμοι απαιτούν τη χαρτογράφηση των στοιχείων σε μια μορφή όπου οι ιδιότητες έχουν τις ιδιαίτερες τιμές για να μπορούν να δουλέψουν με το γενετικό αλγόριθμο. Αυτό γίνεται, αλλά μπορεί να χαθούν πολλές από τις πληροφορίες λεπτομέρειας κατά την εξέταση των μεταβλητών. Η κωδικοποίηση των στοιχείων στην κατηγορική μορφή μπορούν ακούσια να οδηγήσουν σε προκαταλήψεις στα στοιχεία. Υπάρχουν επίσης όρια στο μέγεθος του συνόλου των στοιχείων που μπορεί να αναλυθεί με γενετικούς αλγόριθμους. Για τα πολύ μεγάλα σύνολα στοιχείων, η δειγματοληψία θα είναι απαραίτητη, και οδηγεί σε διαφορετικά αποτελέσματα, στα διαφορετικά "τρεξίματα" του ίδιο σύνολο στοιχείων.

2.10.1 Επίδειξη του γενετικού αλγορίθμου

Η αξία των γενετικών αλγορίθμων είναι στη δυνατότητά τους να εξετάζουν τα σύνθετα σύνολα στοιχείων, όπου υπάρχει ένας αδικαιολόγητα μεγάλος αριθμός συνδυασμών μεταβλητών. Μια από τις πιο συνηθισμένες εφαρμογές της εξόρυξης δεδομένων είναι στις εφαρμογές δανείου. Χρησιμοποιούμε ένα σύνολο δεδομένων εφαρμογής δανείου με τις αντιπροσωπευτικές παρατηρήσεις που δίνονται στο παράρτημα. Ο σκοπός του γενετικού αλγορίθμου είναι σε αυτήν την περίπτωση απλά να προσδιορίσει ένα σύνολο υποψήφιων με τη βέλτιστη λειτουργία ικανότητας. Ο σκοπός είναι να φανεί, πώς ο γενετικός αλγόριθμος θα λειτουργούσε, ανεξάρτητα από το πόσο μεγάλο είναι το σύνολο των στοιχείων. Τα βήματα της εφαρμογής του γενετικού αλγορίθμου αρχίζουν με το να διαχωριστούν τα στοιχεία. Σε αυτήν την περίπτωση, τα στοιχεία έχουν ήδη διαχωριστεί, σε τρεις μεταβλητές, ηλικία, εισόδημα, και κίνδυνος (age, income και risk) κάθε ένας με τρεις πιθανές τιμές. Αυτό παρέχει μια σχετικά απλή περίπτωση δεδομένου στο οποίο υπάρχουν μόνο 27 συνδυασμοί όλων αυτών των μεταβλητών.

Το σύνολο των 20 περιπτώσεων από τον πίνακα 2.7 μπορεί να οργανωθεί μέσα από τη διαίρεση όπως φαίνεται στη μεταβλητή ηλικία (age) έχει τρεις τιμές (νέοι, μέση ηλικία, και γέροι), οι οποίες είναι ονομαστικές. Το μεταβλητό εισόδημα (income) έχει τρεις τιμές επίσης (χαμηλό, μέσο, και υψηλό), οι οποίες είναι τακτικές. Τρεις τιμές του μεταβλητού κινδύνου (risk) (υψηλός, μέσος, χαμηλός) που είναι επίσης τακτικές. Οι εκβάσεις είναι ναι και όχι .

Πίνακας 2.7

Case	Age	Income	Risk	Outcome
1	1	1	1	1
2	1	1	1	1
3	1	1	1	0
4	1	1	1	0
5	1	1	3	1
6	1	2	1	0
7	1	2	1	1
8	1	2	1	1
9	1	2	1	1
10	1	2	3	1
11	1	2	3	1
12	1	3	1	0
13	2	2	2	0
14	2	2	1	1
15	2	2	1	1
16	2	3	1	1
17	2	3	1	1
18	3	3	1	1
19	3	3	1	1
20	3	3	1	1

Πίνακας 2.8

Age	Income	Risk	Not OK	OK	Function
1	1	1	2	2	0.50
1	1	3	0	1	1.00
1	2	1	1	3	0.75
1	2	3	0	2	1.00
1	3	1	1	0	0.00
2	2	1	0	2	1.00
2	2	2	1	0	0.00
2	3	1	0	2	1.00
3	3	1	0	3	1.00

"Η λειτουργία" στους πίνακες που ακολουθούν είναι η πιθανότητα της έγκαιρης επιστροφής του δανείου, που βασίζεται στις πιθανότητες που υπολογίζονται από τον πίνακα 2.7.

Αυτή η λειτουργία θα είναι η βάση για τα επιζόντα γονίδια.

Τρία βήματα περιλαμβάνονται στο γενετικό αλγόριθμο:

1. Αναπαραγωγή
2. Επιλογή
3. Μεταλλαγή

Ο αλγόριθμος ξεκινά από τυχαία παραγομένους γονείς από το σύνολο. Εδώ θα επιλέξουμε αυθαίρετα τέσσερις γονείς. Όσο περισσότεροι γονείς επιλέγονται, τόσο περισσότερο θα διαρκέσει κάθε επανάληψη, αλλά και τόσο πιο λεπτομερής θα είναι ο αλγόριθμος όπως φαίνεται στον πίνακα 2.9.

Θα ζευγαρώσουμε αυτές τις τυχαίες επιλογές για να παραγάγουμε την επόμενη γενεά. Η αναπαραγωγή ολοκληρώνεται με το να διασχίσει τυχαία τις επιλεγμένες μεταβλητές τιμές μεταξύ των γονέων.

Πίνακας 2.9

Case	Age	Income	Risk	Function
3	1	1	1	0.50
13	2	2	2	0.00
12	1	3	1	0.00
7	1	2	1	0.75

Πίνακας 2.10

	Case	Age	Income	Risk	Function
Parent 1	3	1	1	1	0.50
Parent 2	13	2	2	2	0.00
Offspring 1		1	2	1	0.75
Offspring 2		2	1	2	Not observed

Παραδείγματος χάριν, ο πίνακας 2.10 παρουσιάζει τις περιπτώσεις 3 και 13 που ποικίλουν και στις τρεις μεταβλητές. Εάν ο αλγόριθμος επίλεξε τυχαία τη δεύτερη μεταβλητή (εισόδημα) για τη διασταύρωση, οι δύο παραγμένες περιπτώσεις θα ήταν (ηλικία 1, εισόδημα 2, κίνδυνος 1) και (η ηλικία 2, εισόδημα 1, κίνδυνος 2). Αποτελέσματα για τις νέες υποθέσεις μπορεί να ληφθούν από τον αλγόριθμο. Η διασταύρωση μπορεί να επαναλάβει έναν από τους γονείς, και στην πραγματικότητα εάν και οι δύο γονείς έχουν τα ίδια χαρακτηριστικά, η επανάληψη είναι σίγουρη. Εντούτοις, το τρίτο βήμα της μεταλλαγής μπορεί να σπάσει τέτοια αδιέξοδα.

Η επιλογή μπορεί να ακολουθήσει διάφορους κανόνες, αλλά θα επιλέξουμε τις δύο περιπτώσεις με τις καλύτερες τιμές. Σε αυτήν την περίπτωση, παράγουμε μια νέα λύση, η οποία είναι ανώτερη και από τους δύο γονείς. Αυτή η πίεση θα μεταφέρει τις δύο εναλλακτικές λύσεις με τις καλύτερες λειτουργικές τιμές (γονέας 1, απόγονος 1). Για το επόμενο ζευγάρι των τυχαία επιλεγμένων γονέων, μπορούμε τυχαία να επιλέξουμε τη μεταβλητή στη διασταύρωση. Εάν επιλέχτηκε το εισόδημα, τα αποτελέσματα θα ήταν όπως παρουσιάζεται στον πίνακα 2.11.

Εδώ η αναπαραγωγή δεν παρήγαγε τίποτα νέο. Αλλά εάν κάθε μία από τις άλλες δύο μεταβλητές είχε επιλεχτεί, οι απόγονοι θα ήταν κλώνοι των γονιών, πάλι με τίποτα καινούριο. Οι περιπτώσεις όπως αυτή καθιστούν τη μεταλλαγή χρήσιμη. Η μεταλλαγή περιλαμβάνει την τυχαία αλλαγή μιας από τις επιλεγμένες περιπτώσεις. Παραδείγματος χάριν, ο απόγονος 3 θα μπορούσε να αλλοιώσει με την αλλαγή της αξίας για την ηλικία σε κάποια άλλη αξία, όπως 2. Αλγοριθμικά αυτό θα ολοκληρωνόταν τυχαία, και ως προς ποια μεταβλητή ήταν να αλλάξει, καθώς επίσης και σε ποια αξία θα άλλαζε.

Πίνακας 2.11

	Case	Age	Income	Risk	Function
Parent 3	12	1	3	1	0.00
Parent 4	7	1	2	1	0.75
Offspring 3		1	2	1	0.75
Offspring 4		1	3	1	0.00

Πίνακας 2.12

Case	Age	Income	Risk	Function
Parent 1	1	1	1	0.50
Offspring 1	1	2	1	0.75
Parent 4	1	2	1	0.75
Mutation	2	2	1	1.00

Στο τέλος της πρώτης επανάληψης, έχουμε τώρα δύο σύνολα επιζώντων γονέων ανά γενεά, που παρουσιάζεται στον πίνακα 2.12:

Το δεύτερο σύνολο ενεργών γονέων παράγει τις νέες υποθέσεις, επιλέγοντας τυχαία τη μεταβλητή ηλικία για τη διασταύρωση, που παρουσιάζεται στους πίνακες 2.13 και 2.14:

Εδώ έχουμε φθάσει σε ένα αδιέξοδο, με την αντιστροφή των χαρακτηριστικών του απογόνου σε σχέση με τους γονείς. Κατά συνέπεια, αυτή η προσπάθεια έχει φθάσει σε ένα αδιέξοδο όσον αφορά την παραγωγή των νέων λύσεων (εκτός αν εφαρμοστεί η μεταλλαγή). Εδώ λαμβάνουμε δύο περιπτώσεις με ισχυρά χαρακτηριστικά επιβίωσης (μεταλλαγή, απόγονος 5). Κατά συνέπεια ο αλγόριθμος πέτυχε στον προσδιορισμό τριών πολύ καλών λύσεων (η μεταλλαγή, και απόγονος 5). Υπάρχουν επίσης και άλλες καλές λύσεις, αλλά ο σκοπός του γενετικού αλγορίθμου ήταν να προσδιοριστεί οποιαδήποτε καλή λύση. Η επίπτωση είναι ότι οι υποψήφιοι για δάνειο με μέση ηλικία, μέσο εισόδημα, και καλά αποτελέσματα στον κίνδυνο είναι ιδιαίτερα πιθανό να ξεπληρώσουν χωρίς περιπλοκές. Σημειώστε ότι δεν προτείνουμε τους γενετικούς αλγορίθμους για την εξόρυξη δεδομένων ως μέσο πρόβλεψης της αποπληρωμής. Προσπαθούμε να δείξουμε πώς οι γενετικοί αλγόριθμοι μπορούν να παράγουν τις βελτιωμένες λύσεις σε περιοχές με υψηλού βαθμού μη γραμμικότητα, όπου οι παραδοσιακές προσεγγίσεις όπως η γραμμική παλινδρόμηση μπορούν να έχουν δυσκολίες. Οι γενετικοί αλγόριθμοι είναι χρήσιμοι στην εξόρυξη δεδομένων ως συμπληρωματικό εργαλείο.

Πίνακας 2.13

Case	Age	Income	Risk	Function
Parent 1	1	1	1	0.50
Offspring 1	1	2	1	0.75
Offspring 5	1	2	1	0.75
Offspring 6	1	1	1	0.50

Πίνακας 2.14

Case	Age	Income	Risk	Function
Parent 4	1	2	1	0.75
Mutation	2	2	1	1.00
Offspring 5	2	2	1	1.00
Offspring 6	1	2	1	0.75

2.10.2 Εφαρμογή των γενετικών αλγορίθμων στην εξόρυξη δεδομένων

Οι γενετικοί αλγόριθμοι έχουν εφαρμοστεί στην εξόρυξη δεδομένων με δύο τρόπους. Η εξωτερική υποστήριξη είναι μέσω της αξιολόγησης ή της βελτιστοποίησης κάποιας παραμέτρου για κάποιο άλλο σύστημα εκμάθησης, συχνά υβριδικά συστήματα που χρησιμοποιούν άλλα εργαλεία εξόρυξης δεδομένων όπως τα δέντρα συγκέντρωσης ή απόφασης. Από αυτή την άποψη, οι γενετικοί αλγόριθμοι βοηθούν άλλα εργαλεία εξόρυξης δεδομένων να λειτουργούν αποτελεσματικότερα. Οι γενετικοί αλγόριθμοι μπορούν επίσης να εφαρμοστούν άμεσα στην ανάλυση, όπου ο γενετικός αλγόριθμος παράγει τις περιγραφές, συνήθως ως κανόνες απόφασης ή δέντρα απόφασης. Πολλές εφαρμογές των γενετικών αλγορίθμων μέσα στην εξόρυξη δεδομένων έχουν εφαρμοστεί έξω από την επιχείρηση.

Συγκεκριμένα παραδείγματα περιλαμβάνουν την εξόρυξη δεδομένων για ιατρικούς λόγους και την ανίχνευση παρείσφρησης σε δίκτυα υπολογιστών. Στις επιχειρήσεις, οι γενετικοί αλγόριθμοι έχουν εφαρμοστεί στην κατάτμηση πελατών, την πιστωτική φερεγγυότητα , και την οικονομική επιλογή ασφάλειας.

Οι γενετικοί αλγόριθμοι μπορούν να είναι πολύ χρήσιμοι μέσα σε μια ανάλυση εξόρυξης δεδομένων που έχει να εξετάσει περισσότερες ιδιότητες και πολλές περισσότερες παρατηρήσεις. Εντούτοις, η εφαρμογή των γενετικών αλγορίθμων απαιτεί την έκφραση των στοιχείων σε ξεχωριστές εκβάσεις, με μια υπολογίσιμη λειτουργική αξία στην οποία θα βασίσει την επιλογή.

Οι γενετικοί αλγόριθμοι λειτουργούν στα διακριτά στοιχεία με συστηματική αναζήτηση των καλύτερων (ή ενδεχομένως βέλτιστων) συνδυασμών μεταβλητών τιμών. Μπορούν να είναι πολύ αποδοτικοί στα προβλήματα με σύνθετες αλληλεπιδράσεις, ειδικά όταν περιλαμβάνονται μη γραμμικές λειτουργίες. Οι γενετικοί αλγόριθμοι είναι πολύτιμοι στην εξόρυξη δεδομένων λόγω της δυνατότητάς τους να εξετάσουν τις ανακριβείς και αντιφατικές πληροφορίες. Εφαρμόζονται συνήθως από κοινού με άλλες τεχνικές εξόρυξης δεδομένων. Μπορούν να χρησιμοποιηθούν για να ενισχύσουν την αποδοτικότητα άλλων μεθόδων, ή μπορούν να εφαρμοστούν αμεσότερα. (David L. Olson & Dursun Delen, 2008)

2.11 Προετοιμασία των δεδομένων για τη διαδικασία της εξόρυξης δεδομένων

2.11.1 Προετοιμασία των δεδομένων για όλα τα εργαλεία εξόρυξης δεδομένων

Πολύ συχνά απαιτείται να γίνει προετοιμασία των δεδομένων, προκειμένου να εισέλθουν στην διαδικασία της εξόρυξης δεδομένων. Μάλιστα μπορεί να δαπανηθεί περισσότερος χρόνος κατά την προετοιμασία των δεδομένων, παρά κατά την διαδικασία εξόρυξης των δεδομένων. Τα κυριότερα βήματα, προκειμένου τα δεδομένα να αποκτήσουν την απαιτούμενη μορφή, είναι τα εξής (Kimball 1997):

- Η διόρθωση μη συνεπούς μορφής δεδομένων και η διόρθωση μη συμβατικής κωδικοποίησης των δεδομένων, συντομογραφιών και σημείων στίξης
- Η αφαίρεση ανεπιθύμητων ή περιττών πεδίων. Τα δεδομένα μπορεί να περιέχουν πεδία άνευ σημασίας για την ανάλυση που θέλουμε να κάνουμε. Τα εργαλεία της εξόρυξης δεδομένων ενδέχεται να ερμηνεύσουν αυτά τα πεδία ως μετρήσεις ή μεγέθη, ειδικά αν είναι αριθμητικά , και μπορεί να κάνουν κύκλους προσπαθώντας να βρουν συγκεκριμένα μοτίβα με αυτά τα πεδία, ή προσπαθώντας να συσχετίσουν αυτά τα πεδία με πραγματικά δεδομένα
- Η ερμηνεία των κωδικών σε κείμενο. Η κλασσική μορφή «καθαρισμού» των δεδομένων περιλαμβάνει τη βελτίωση ή την αντικατάσταση αινιγματικών κωδικών με κείμενο, γραμμένο με αναγνωρίσιμες λέξεις.
- Ο συνδυασμός των δεδομένων που προέρχονται από διάφορες πηγές, όπως τα δεδομένα των πελατών, σε μια κοινή βάση.
- Η εύρεση των πεδίων που έχουν χρησιμοποιηθεί για παραπάνω από έναν σκοπό. Ένας καλός τρόπος για να βρεθούν αυτά τα αρχεία είναι η καταμέτρηση, και ίσως και η δημιουργία μιας λίστας όλων των διαφορετικών τιμών-χρήσεων που υπάρχουν σε ένα πεδίο.
- Ο έλεγχος για τυχόν μη φυσιολογικά, εκτός ορίων η αδύνατα στοιχεία. Κάποια μετρήσιμα στοιχεία μπορεί να είναι σωστά, αλλά εξαιρετικά ασυνήθιστα. Τέτοιου είδους δεδομένα, είναι προτιμότερο να μαρκαριστούν με μια ειδική σήμανση, έτσι ώστε να μπορούμε να τα συμπεριλάβουμε ή να τα εξαιρέσουμε από την ανάλυσή μας, ανάλογα με την περίπτωση.

- Ο έλεγχος για τυχόν τιμές που λείπουν, ή εάν αυτές έχουν αντικατασταθεί από κάποιον προεπιλεγμένο αριθμό.
- Η εφαρμογή ομοιόμορφης μεταχείρισης σε μηδενικές τιμές. Οι μηδενικές τιμές ενδέχεται να δυσκολέψουν την λειτουργία του εργαλείου εξόρυξης δεδομένων. Σε πολλές περιπτώσεις, η μηδενική αξία αντιπροσωπεύεται από κάποια άλλη ιδιαίτερη τιμή. Πολλές φορές η τιμή-1 θεωρείται ότι αντιπροσωπεύει τις μηδενικές αξίες.
- Η ταξινόμηση μεμονωμένων αρχείων δεδομένων σύμφωνα με ένα από τα συγκεντρωτικά του μεγέθη. Σε ορισμένες περιπτώσεις, μπορεί να είναι επιθυμητός ο εντοπισμός της πώλησης ενός πολύ συγκεκριμένου προϊόντος, όπως ένα ρούχο σε ένα συγκεκριμένο χρώμα και μέγεθος, και από ένα συγκεκριμένο υλικό.

2.11.2 Προετοιμασία των δεδομένων ανάλογα με το χρησιμοποιούμενο εργαλείο εξόρυξης δεδομένων.

Ανάλογα με το χρησιμοποιούμενο εργαλείο εξόρυξης δεδομένων, μπορεί να χρειαστεί να γίνουν κάποιες επιπλέον μετατροπές στα δεδομένα , πέρα των προηγουμένων, και αυτές είναι οι ακόλουθες:

- Ο διαχωρισμός των πρωτογενών δεδομένων εισόδου σε τρείς ομάδες. Η πρώτη ομάδα δεδομένων χρησιμοποιείται για την κατάρτιση του εργαλείου εξόρυξης δεδομένων. Ένα εργαλείο συσταδοποίησης (clustering tool), ένα εργαλείο νευρωνικών δικτύων (neural network tools), ή ένα εργαλείο δέντρου αποφάσεων (decision tree tool) απορροφά την πρώτη σειρά στοιχείων και ορίζει τις παραμέτρους, από τις οποίες μπορούν να γίνουν οι μελλοντικές ταξινομήσεις και οι προβλέψεις. Το δεύτερο σύνολο δεδομένων, που χρησιμοποιείται στη συνέχεια, ελέγχει αυτές τις παραμέτρους για να δει πόσο καλά αποδίδει το μοντέλο. Όταν το εργαλείο εξόρυξης δεδομένων έχει ρυθμιστεί σωστά στο πρώτο και στο δεύτερο σύνολο δεδομένων, εφαρμόζεται στη συνέχεια στην τρίτη σειρά αξιολόγησης των δεδομένων, όπου τα συμπλέγματα, οι ταξινομήσεις και οι προβλέψεις που προέρχονται από το εργαλείο είναι πλήρως αξιόπιστες και μπορούμε να τις χρησιμοποιήσουμε.
- Η προσθήκη υπολογισμένων πεδίων ως εισροές ή ως στόχοι. Για παράδειγμα, ένα υπολογιζόμενο πεδίο, όπως τα κέρδη ή η ικανοποίηση των πελατών, που αντιπροσωπεύει την αξία ενός συνόλου συναλλαγών των πελατών, μπορεί να χρειαστεί να τεθεί ως στόχος για να επιλέξει το εργαλείο εξόρυξης δεδομένων τους

ποιο κερδοφόρους πελάτες ή για να επιλέξει τη συμπεριφορά που θέλουμε να ενθαρρύνουμε.

- Η διάταξη των συνεχών τιμών σε κλίμακες. Μερικά εργαλεία εξόρυξης δεδομένων, όπως τα δέντρα αποφάσεων, ενθαρρύνουν τη διάταξη αυτή σε διακριτές κλίμακες.
- Η εξομάλυνση των τιμών μεταξύ 0 και 1. Τα εργαλεία νευρωνικών δικτύων συνήθως απαιτούν όλες οι αριθμητικές τιμές να αντιστοιχίζονται σε μια σειρά από το μηδέν έως το ένα.
- Η μετατροπή των κειμένων σε αριθμητικές τιμές. Μερικά εργαλεία εξόρυξης δεδομένων μπορούν να λειτουργήσουν μόνο με αριθμητικά δεδομένα εισόδου. Σε αυτές τις περιπτώσεις, οι διακριτές τιμές κείμενου θα πρέπει να αντικατασταθούν από ειδικούς κωδικούς, όπως για παράδειγμα η αντικατάσταση της περιοχής του κάθε πελάτη με τον αντίστοιχο ταχυδρομικό κώδικα.

Κεφάλαιο 3

Εξόρυξη Δεδομένων στα συστήματα CRM

3.1 Εξόρυξη δεδομένων και CRM

Σε αυτό το κεφάλαιο θα ασχοληθούμε με την ανάπτυξη μιας βάσης δεδομένων σχετικής με τον πελάτη σε μία εφαρμογή CRM. Ο καθορισμός των αναγκών και των στόχων της επιχείρησης, η σωστή δημιουργία και η διατήρηση της βάσης δεδομένων είναι καθοριστική και κρίσιμη για την επιτυχία του συστήματος CRM. Στη συνέχεια θα εξηγήσουμε το πώς χρησιμοποιείται η εξόρυξη δεδομένων (data-mining) στις εφαρμογές CRM, και τι προσφέρει στο χρήστη.

3.1.1 Ανάπτυξη μιας βάσης δεδομένων σχετικής με τον πελάτη

Οι περισσότερες βάσεις δεδομένων μοιράζονται μια κοινή δομή αρχείων, στοιχείων και τομέων (πίνακες, σειρές, στήλες). Τα αρχεία (πίνακες) φυλάσσουν τις πληροφορίες για ένα ενιαίο θέμα όπως οι πελάτες, τα προϊόντα, οι συναλλαγές ή τα αιτήματα υπηρεσιών. Κάθε αρχείο (πίνακας) περιέχει διάφορα αρχεία (σειρές). Κάθε αρχείο (σειρά) περιλαμβάνει διάφορα στοιχεία των δεδομένων. Αυτά τα στοιχεία τακτοποιούνται στα κοινά σύνολα τομέων (στήλες) στον πίνακα. Η σύγχρονη βάση δεδομένων σχετική με τον πελάτη επομένως μοιάζει με υπολογισμό σε λογιστικό φύλλο (spreadsheet). Υπάρχουν έξι σημαντικά βήματα στην οικοδόμηση μιας βάσης δεδομένων σχετικής με τον πελάτη, όπως φαίνεται παρακάτω.

1. Καθορισμός των λειτουργιών της βάσης δεδομένων
2. Καθορισμός των αναγκών για πληροφόρηση
3. Προσδιορισμός των πηγών των πληροφοριών
4. Επιλογή της πλατφόρμας τεχνολογίας και του υλικού βάσης δεδομένων
5. Δημιουργία της βάσης δεδομένων
6. Διατήρηση της βάσης δεδομένων

1. Καθορισμός των λειτουργιών της βάσης δεδομένων

Οι βάσεις δεδομένων υποστηρίζουν τις τέσσερις μορφές του CRM - στρατηγικό, λειτουργικό, αναλυτικό και συνεργατικό. Το στρατηγικό CRM χρειάζεται τα στοιχεία για τις

αγορές, τις προσφορές αγοράς, τους πελάτες, τα κανάλια, τους ανταγωνιστές, την απόδοση και τη δυνατότητα να είναι σε θέση να προσδιορίσει ποιοι είναι οι πελάτες-στόχοι για απόκτηση, τη διατήρηση και την ανάπτυξη πελατών, και τι να τους προσφέρει. Οι εφαρμογές του συνεργατικού CRM χρησιμοποιούν γενικά τα λειτουργικά και αναλυτικά στοιχεία όπως περιγράφονται παρακάτω, έτσι ώστε οι συνεργάτες της επιχείρησης στα κανάλια διανομής, να μπορούν να ευθυγραμμίσουν τις προσπάθειές τους για εξυπηρέτηση των τελικών πελατών. Τα σχετικά με τον πελάτη στοιχειά είναι απαραίτητα για τους σκοπούς του λειτουργικού και του αναλυτικού CRM. Το λειτουργικό CRM χρησιμοποιεί τα σχετικά με τον πελάτη στοιχεία που βοηθούν στην καθημερινή λειτουργία της επιχείρησης.

2. Καθορισμός των αναγκών για πληροφόρηση

Οι άνθρωποι που μπορούν να απαντήσουν καλύτερα στην ερώτηση «ποιες πληροφορίες χρειάζονται;» είναι εκείνοι που αλληλεπιδρούν ή επικοινωνούν με τους πελάτες για πωλήσεις, σκοπούς του μάρκετινγκ και των υπηρεσιών, και εκείνοι που πρέπει να πάρουν στρατηγικές αποφάσεις CRM. Κάποιος που προγραμματίζει μια εκστρατεία μέσω ηλεκτρονικού ταχυδρομείου θα θέλει να ξέρει τα ποσοστά open and click-through και τα ποσοστά click-to-open (CTOR) από προηγούμενες e-εκστρατείες, που κατανέμονται ανά αγορά, προσφορά και εκτέλεση. Θα ήθελε επίσης να ξέρει τις διευθύνσεις ηλεκτρονικού ταχυδρομείου, τις προτιμήσεις για ηλεκτρονικό ταχυδρομείο (HTML ή απλό κείμενο), και προτιμημένος χαιρετισμός (όνομα;, Κος;, Κα;) Οι λειτουργικές και αναλυτικές ανάγκες όπως αυτές βοηθούν στο καθορισμό του περιεχόμενο των βάσεων δεδομένων σχετικών με τους πελάτες.

Ανώτερα στελέχη που αναθεωρούν τις στρατηγικές αποφάσεις του CRM της επιχείρησής τους θα επιθυμούσαν ένα απολύτως διαφορετικό σύνολο πληροφοριών. Μπορεί να θέλουν να ξέρουν τα ακόλουθα. Πώς χωρίζεται η αγορά; Ποιοι είναι οι τρέχοντες πελάτες μας; Τι αγοράζουν; Από ποιους άλλους αγοράζουν; Ποιες είναι οι απαιτήσεις, οι προσδοκίες και οι προτιμήσεις των πελατών μας σε όλα τα μέρη της πρότασης μας, συμπεριλαμβανομένου του προϊόντος, της υπηρεσίας, του καναλιού και της επικοινωνίας;

Με την εμφάνιση των CRM εφαρμογών, ένα μεγάλο μέρος του σχεδιασμού της βάσης δεδομένων έχει γίνει από τους προμηθευτές του λογισμικού. Η διαθεσιμότητα σε εφαρμογές CRM εξειδικευμένες για επιχειρήσεις, με τα αντίστοιχα εξειδικευμένα πρότυπα στοιχείων τους, επιτρέπουν μια πολύ πιο στενή ταυτοποίηση με τις ανάγκες της επιχείρησης για δεδομένα.

3. Προσδιορισμός των πηγών των πληροφοριών

Οι πληροφορίες για τις βάσεις δεδομένων σχετικές με τον πελάτη μπορούν να πηγάζουν εσωτερικά ή εξωτερικά. Πριν από την οικοδόμηση της βάσης δεδομένων είναι απαραίτητο η επιχείρηση να ελέγξει, για να ανακαλύψει ποια στοιχεία είναι διαθέσιμα. Τα εσωτερικά στοιχεία είναι τα θεμέλια για τα περισσότερα CRM προγράμματα, εν τούτοις το πλήθος των πληροφοριών που είναι διαθέσιμα για τους πελάτες εξαρτάται από το βαθμό επαφής που έχει η επιχείρηση με τον πελάτη. Μερικές επιχειρήσεις πωλούν μέσω συνεργατών, πρακτορειών και διανομέων και έχουν λίγη γνώση για την αλυσίδα απαιτήσεων πέρα από την άμεση επαφή τους.

Τα εσωτερικά στοιχεία μπορούν να βρεθούν στις διάφορες λειτουργικές περιοχές

- Το μάρκετινγκ μπορεί να έχει τα στοιχεία όσον αφορά το μέγεθος της αγοράς, την κατάτμηση της αγοράς, τα προφίλ των πελατών, τα κανάλια αποκτήσεων πελατών, τα αρχεία των εκστρατειών μάρκετινγκ , τις εγγραφές προϊόντων και τα αιτήματα για τις πληροφορίες προϊόντων.
- Οι πωλήσεις μπορεί να έχουν τα αρχεία για το ιστορικό αγορών των πελατών συμπεριλαμβανομένου τη συχνότητα και τη νομισματική αξία, τα ονόματα των αγοραστών και στοιχεία επαφής, αριθμός απολογισμού, κώδικα SIC, σημαντικά κριτήρια αγοράς, εμπορικοί όροι όπως οι εκπτώσεις και οι περίοδοι πληρωμής, πιθανοί πελάτες (προοπτικές), απαντήσεις στις προτάσεις, προϊόντα ανταγωνιστών και τιμολόγηση, και απαιτήσεις και προτιμήσεις των πελατών.
- Η εξυπηρέτηση πελατών μπορεί να έχει τα αρχεία του ιστορικού υπηρεσιών, απαιτήσεις των υπηρεσιών , επίπεδα ικανοποίησης πελατών, καταγγελίες πελατών, επιλυμένα και εκκρεμή ζητήματα, έρευνες, και προγράμματα πίστης , ιδιότητα μέλους και θέση.
- Το λογιστήριο μπορεί να έχει τα στοιχεία όσον αφορά τις πιστωτικές εκτιμήσεις, προϋπολογισμοί και ιστορικό πληρωμών.
- Ο υπεύθυνος για το διαδίκτυο μπορεί να έχει τα click-stream δεδομένα.

4. Επιλογή της πλατφόρμας τεχνολογίας και του υλικού βάσης δεδομένων

1. Ιεραρχική
2. Δικτυακή
3. Σχεσιακή

Οι ιεραρχικές και οι βάσεις δεδομένων δικτύων ήταν η πιο κοινή μορφή μεταξύ της δεκαετίας του '60 και της δεκαετίας του '80. Η ιεραρχική βάση δεδομένων είναι η παλαιότερη μορφή και δεν ταιριάζει στις περισσότερες εφαρμογές CRM. Μπορείτε να φανταστείτε το ιεραρχικό πρότυπο ως διάγραμμα οργάνωσης ή ένα οικογενειακό δέντρο, στα οποία ένα παιδί μπορεί να έχει μόνο έναν γονέα, αλλά ένας γονέας μπορεί να έχει πολλά παιδιά. Ο μόνος τρόπος να αποκτηθεί η πρόσβαση στα χαμηλότερα επίπεδα είναι να αρχίσεις από την κορυφή και προς τα κάτω. Όταν το στοιχείο αποθηκεύεται με το ιεραρχικό σχήμα, μπορεί κάποιος να καταλήξει να δουλεύει σε διάφορα στρώματα των υψηλότερου επιπέδου στοιχείων πριν φτάσει στα στοιχεία που χρειάζεται. Οι βάσεις δεδομένων προϊόντων είναι γενικά ιεραρχικές. Μια σημαντική κατηγορία προϊόντων θα υποδιαιρεθεί επανειλημμένα μέχρι όλες οι μορφές του προϊόντος έχουν καταγραφεί.

5. Δημιουργία της βάσης δεδομένων

Έχοντας αποφασίσει ποιες πληροφορίες, ποια βάση δεδομένων και ποιες απαιτήσεις υλικού απαιτούνται, ο επόμενος στόχος είναι να ληφθούν τα στοιχεία και να εισαχθούν στη βάση δεδομένων. Οι εφαρμογές CRM χρειάζονται τα στοιχεία που είναι κατάλληλα ακριβή. Χρησιμοποιούμε το "κατάλληλα" επειδή το επίπεδο ακρίβειας εξαρτάται από τη λειτουργία της βάσης δεδομένων. Οι εφαρμογές του λειτουργικού CRM γενικά χρειάζονται ακριβέστερα και πιο σύγχρονα στοιχεία από ότι οι αναλυτικές εφαρμογές.

Μπορεί να έχετε δοκιμάσει προσωπικά τα αποτελέσματα των κακής ποιότητας στοιχείων. Ίσως έχετε λάβει μια ταχυδρομημένη πρόσκληση για να γίνετε χορηγός σε μια φιλανθρωπία, στην οποία δίνετε ήδη άμεσα από το μισθό σας, μέσω του τραπεζικού λογαριασμού σας. Αυτό θα μπορούσε να έχει συμβεί όταν ένας κατάλογος έρευνας που έχει αγοραστεί από μια φιλανθρωπική οργάνωση δεν ελέγχθηκε σε σχέση με τους τρέχοντες καταλόγους των δωρητών. Ίσως χρησιμοποιούν το Κα. ενώ το σωστό για την περίπτωση σας είναι Δεσποινίς. Αυτό γίνεται επειδή η επιχείρηση είτε δεν έχει λάβει είτε δεν έχει ενεργήσει είτε δεν έχει ελέγξει τις προτιμήσεις επικοινωνίας σας.

Ένα από τα μεγαλύτερα ζητήματα με τα στοιχεία πελατών δεν είναι τόσο πολύ τα ανακριβές στοιχεία όσο οι απώλειες των στοιχείων. Πολλές οργανώσεις δυσκολεύονται να αποκτήσουν ακόμα και βασικά στοιχεία πελατών, όπως οι διευθύνσεις ηλεκτρονικού ταχυδρομείου και οι προτιμήσεις. Τα βασικά βήματα για να εξασφαλίσουμε ότι η βάση δεδομένων αποτελείται από κατάλληλα και ακριβή στοιχεία είναι τα ακόλουθα:

- Έλεγχος της πηγής των στοιχείων
- Έλεγχος των στοιχείων

- Επικύρωση των στοιχείων
- Αντιγραφή των στοιχείων ξανά
- Συγχώνευση και ξεκαθάριση των στοιχείων από δύο ή περισσότερες πηγές.

6. Διατήρηση της βάσης δεδομένων

Οι βάσεις δεδομένων πελατών πρέπει να ενημερώνονται για να συνεχίζουν να είναι χρήσιμες. Προσέξτε τις ακόλουθες στατιστικές:

- 19% των διευθυντών διοίκησης αλλάζουν εταιρία κάθε χρόνο
- 8% των επιχειρήσεων αλλάζουν έδρα κάθε χρόνο
- Στο Ηνωμένο Βασίλειο, το 5% των ταχυδρομικών κωδίκων αλλάζουν κάθε χρόνο
- Στις δυτικές οικονομίες το 1,2% του πληθυσμού πεθαίνει κάθε χρόνο
- Στις ΗΠΑ, πάνω από 40 εκατομμύρια άνθρωποι αλλάζουν διεύθυνση κάθε χρόνο.

Δεν παίρνει πολύ για να υποβιβαστούν οι βάσεις δεδομένων. Οι επιχειρήσεις μπορούν να διατηρήσουν την ακεραιότητα των στοιχείων με διάφορους τρόπους.

1. Εξασφαλίζοντας ότι τα στοιχεία από όλες τις νέες συναλλαγές, εκστρατείες και επικοινωνίες περιλαμβάνονται στη βάση δεδομένων άμεσα. Θα χρειαστεί να αναπτυχθούν κανόνες και να εξασφαλιστεί ότι εφαρμόζονται.

2. Αντιγράφοντας ξανά τις βάσεις δεδομένων τακτικά.

3 Έλεγχος ενός υποσυνόλου των στοιχείων κάθε χρόνο. Μέτρηση του ποσού υποβάθμισης. Προσδιορισμός της πηγής υποβάθμισης: είναι μια πηγή ενός ιδιαίτερου στοιχείου ή ο τομέας;

4. Ξεκαθάρισμα των πελατών που είναι ανενεργοί για μια ορισμένη χρονική περίοδο. Για τα προϊόντα που αγοράζονται πιο συχνά, η περίοδος παύσης αγορών να είναι έξι μήνες ή λιγότεροι. Για τα προϊόντα με έναν πιο μακροχρόνιο κύκλο αγορών επανάληψης, η περίοδος θα είναι πιο μεγάλη. Δεν είναι πάντα σαφές ποια είναι η κατάλληλη περίοδος λήθαργου . Μερικοί χρήστες πιστωτικών καρτών, παραδείγματος χάριν, μπορούν να έχουν διαφορετικές κάρτες για διαφορετικά νομίσματα. Η αδράνεια για ένα έτος δείχνει ότι ο ιδιοκτήτης δεν έχει ταξιδέψει σε μια χώρα το προηγούμενο έτος. Ο ιδιοκτήτης μπορεί να κάνει διάφορα ταξίδια στο ερχόμενο έτος.

5. "Τροφοδοσία" της βάση δεδομένων. Κάθε φορά που υπάρχει μια επαφή με τον πελάτη υπάρχει η ευκαιρία να προστεθεί νέος ή να ελεγχθούν τα υπάρχοντα στοιχεία.

6. Κάνοντας τους πελάτες να ενημερώσουν τα αρχεία τους. Όταν πελάτες του Amazon αγοράζουν on-line, χρειάζεται να επικυρώσουν ή να ανανεώσουν τα στοιχεία επικοινωνίας και αποστολής.

Επιθυμητές ιδιότητες στοιχείων

Η διατήρηση της βάσης δεδομένων σημαίνει ότι οι χρήστες θα μπορούν να καλύψουν την ανάγκη τους για ακριβή και σχετικά στοιχεία. Ακρίβεια και σχετικότητα είναι δύο από τις έξι επιθυμητές ιδιότητες στοιχείων - τα στοιχεία πρέπει να είστε διαμοιράσιμα (shareable), μεταφερόμενα (transportable), ακριβή (accurate), σχετικά (relevant), έγκαιρα (timely) και ασφαλή (secure) [STARTS].

Τα στοιχεία πρέπει να είναι διαμοιράσιμα επειδή διάφοροι χρήστες μπορεί να απαιτήσουν την πρόσβαση στα ίδια στοιχεία συγχρόνως. Για παράδειγμα, οι πληροφορίες για το προφίλ των πελατών που έχουν αγοράσει την ετήσια ασφάλεια ταξιδιού μπορεί να χρειαστεί να γίνουν διαθέσιμες στους πράκτορες εξυπηρέτησης πελατών σε αρκετές γεωγραφικές θέσεις ταυτόχρονα δεδομένου ότι εξετάζουν τις έρευνες πελατών σε μια διαφημιστική εκστρατεία.

Τα στοιχεία πρέπει να είναι μεταφερόμενα από τη θέση αποθήκευσης, στο χρήστη. Τα στοιχεία πρέπει να είναι διαθέσιμα οπουδήποτε και όποτε ο χρήστης τα αναζητήσει.

Ο χρήστης μπορεί να είναι αντιπρόσωπος εξυπηρέτησης πελατών, οδηγός-διανομέας καθοδόν σε μια παραλαβή, ένας ανεξάρτητος σύμβουλος υποθηκών ή ένας πωλητής μπροστά σε μια προοπτική πώλησης. Οι πολυεθνικές εταιρίες σήμερα με πελάτες σε όλο τον κόσμο, χαρτοφυλάκια προϊόντων σε αρκετές κατηγορίες και πολλαπλάσιες διαδρομές στην αγορά αντιμετωπίζουν προβλήματα μεταφοράς στοιχείων. Οι ηλεκτρονικές βάσεις δεδομένων πελατών είναι ουσιαστικές για τις σημερινές επιχειρήσεις, μαζί με τη διευκόλυνση των τεχνολογιών, όπως τον συγχρονισμό στοιχείων, ασύρματες επικοινωνίες και φυλλομετρητές ιστοσελίδων (web browsers) που κάνουν τα στοιχεία πλήρως μεταφερόμενα.

Η ακρίβεια των στοιχείων είναι ένα ενοχλητικό ζήτημα. Σε έναν ιδανικό κόσμο θα ήταν θαυμάσιο αν είχε κάποιος 100% ακριβή στοιχεία. Αλλά η ακρίβεια των στοιχείων φέρνει υψηλές δαπάνες. Τα στοιχεία λαμβάνονται, εισάγονται, ενσωματώνονται και αναλύονται σε διάφορες στιγμές. Οποιεσδήποτε ή όλες αυτές οι διαδικασίες μπορούν να είναι πηγή για κάποια ανακρίβεια. Τα λάθη πληκτρολόγησης μπορούν να προκαλέσουν τα λάθη στο σημείο εισαγωγής των δεδομένων. Οι ακατάλληλες αναλυτικές διαδικασίες μπορούν να οδηγήσουν σε λάθος συμπεράσματα. Στο CRM, η ανακρίβεια των στοιχείων μπορεί να οδηγήσει σε αδικαιολόγητα λάθη μέσα σε εκστρατείες μάρκετινγκ, ακατάλληλη έρευνα από

τους πωλητές και γενικά κακή εμπειρία για τους πελάτες. Διαβρώνει επίσης την εμπιστοσύνη στο σύστημα CRM, μειώνοντας κατά συνέπεια τη χρήση του. Αυτό οδηγεί στην περαιτέρω υποβάθμιση της ποιότητα των στοιχείων. Για να αντιμετωπιστεί αυτό, οι όγκοι χρήσης και η ποιότητα των στοιχείων πρέπει να ελέγχονται. Τα στοιχεία πρέπει να εισαχθούν από την πηγή παρά από δεύτερο χέρι. Ο ανάγκες του χρήστη πρέπει να ρυθμιστούν. Οι ποιοτικές διαδικασίες στοιχείων όπως η ανάγκη για επαναντιγραφή πρέπει να εισαχθούν. Το βιβλιοπωλείο WH Smith προσδίδει υψηλούς ρυθμούς ανταπόκρισης του CRM για το άμεσο μάρκετινγκ στην ακρίβεια της βάσης δεδομένων τους. Παραδείγματος χάριν, μια προσφορά για το βιβλίο "Delia Smith's How to Cook book" πέτυχε ένα ποσοστό ανταπόκρισης 8%, σημαντικά περισσότερο από ήταν προτού εφαρμοστεί το πρόγραμμα ποιότητας των στοιχείων τους.

Το σχετικό στοιχείο είναι σχετικό για έναν δεδομένο σκοπό. Για να ελέγξεις την πιστωτική αξία ενός πελάτη, χρειάζεσαι το ιστορικό των συναλλαγής και πληρωμών τους, και την τρέχουσα θέση απασχόλησης και εισοδήματος. Για να βρεις τους πελάτες που παρουσιάζουν προοπτικές για μια σταυροειδείς εκστρατεία, χρειάζεστε τα ποσοστά ροπής για αγορά. Στο σχεδιασμό ενός συστήματος διαχείρισης στοιχείων για να υποστηρίξει μια στρατηγική CRM, η σχετικότητα είναι ένα σημαντικό ζήτημα. Πρέπει να ξέρουμε ποιες αποφάσεις θα παρθούν και ποιες πληροφορίες απαιτούνται για να επιτρέψουν σε αυτές να γίνουν σωστά.

Τα έγκαιρα στοιχεία είναι στοιχεία που είναι διαθέσιμα όπως και όταν είναι απαραίτητα. Στοιχεία που ανακτήθηκαν αφότου λήφθηκε η απόφαση, δεν είναι χρήσιμα. Εξίσου, οι υπεύθυνοι για τις αποφάσεις δεν θέλουν να "φορτώνονται" με τα στοιχεία προτού να γίνει αισθητή η ανάγκη. Οι ταμίες στις τράπεζες χρειάζονται τις πληροφορίες για τη "ροπή για αγορά" διαθέσιμες σε αυτούς όταν ένας πελάτης εξυπηρετείται.

Η ασφάλεια των στοιχείων είναι ένα πολύ σημαντικό ζήτημα για τις περισσότερες επιχειρήσεις. Τα στοιχεία, ιδιαίτερα τα στοιχεία για τους πελάτες, είναι ένας σημαντικός πόρος και μια πηγή ανταγωνιστικού πλεονεκτήματος. Αποτελεί τη βάση για την παράδοση καλύτερων λύσεων στους πελάτες. Οι επιχειρήσεις πρέπει να προστατεύσουν τα στοιχεία τους από την απώλεια, την δολιοφθορά και την κλοπή. Πολλές επιχειρήσεις κάνουν back-up τακτικά τα στοιχεία τους. Η ασφάλεια ενισχύεται μέσω των φυσικών και ηλεκτρονικών εμποδίων όπως τα firewalls. Η διαχείριση της ασφάλειας στοιχείων σε ένα περιβάλλον συνεργατών είναι πρόκληση, μιας και είναι σαφές ότι οι ανταγωνιστικοί συνεργάτες δεν βλέπουν τις πωλήσεις και πληροφορίες ευκαιριών του άλλου, παρά την υπογραφή στο ίδιο σύστημα CRM μέσω της ίδιας πύλης. (Francis Buttle, 2009)

3.1.2 Εξόρυξη δεδομένων

Σύμφωνα με τον Francis Buttle (2009) στο CRM, η εξόρυξη δεδομένων μπορεί να καθοριστεί ως εξής: Η εξόρυξη δεδομένων είναι η εφαρμογή της περιγραφής και ανάλυσης προβλέψεων (descriptive and predictive analysis) για να υποστηρίξει το μάρκετινγκ, τις πωλήσεις και τις λειτουργίες εξυπηρέτησης.

Αν και η εξόρυξη δεδομένων μπορεί να εκτελεστεί στις λειτουργικές βάσεις δεδομένων, εφαρμόζεται συχνότερα στα σταθερότερα σύνολα δεδομένων που κρατούνται σε αποθήκες δεδομένων (data warehouses). Υψηλότερες ταχύτητες επεξεργασίας, μειωμένες δαπάνες αποθήκευσης και καλύτερα πακέτα λογισμικού έχουν καταστήσει την εξόρυξη δεδομένων ελκυστικότερη και πιο οικονομική.

Η εξόρυξη δεδομένων μπορεί να δώσει απαντήσεις στις ερωτήσεις που είναι σημαντικές για τους στρατηγικούς και λειτουργικούς σκοπούς του CRM.

Όπως για παράδειγμα:

- Πώς μπορούν η αγορά και η βάση πελατών μας να διασταυρωθούν;
- Ποιοι πελάτες είναι πιο πολύτιμοι;
- Ποιοι πελάτες προσφέρουν την περισσότερη δυνατότητα για το μέλλον;
- Ποιοι τύποι πελατών αγοράζουν τα προϊόντα μας; Ή δεν αγοράζουν;
- Υπάρχουν οποιαδήποτε σχέδια αγοραστικής συμπεριφοράς στη βάση πελατών μας;
- Θα έπρεπε να χρεώνουμε την ίδια τιμή σε όλα αυτά τα τμήματα;
- Ποιο είναι το προφίλ των πελατών που προκαθορίζουν την πληρωμή;
- Ποιες είναι οι δαπάνες για την απόκτηση πελατών;
- Ποια είδη πελατών θα πρέπει να στοχεύσει για απόκτηση η επιχείρηση;
- Ποιες προσφορές θα έπρεπε να υποβληθούν σε συγκεκριμένες ομάδες πελατών για να αυξηθεί η αξία τους;
- Ποιους πελάτες θα πρέπει να επιδιώξει να διατηρήσει;
- Ποια τακτική διατήρησης δουλεύει καλά;

Η εξόρυξη δεδομένων βοηθάει το CRM με διάφορους τρόπους. Μπορεί να βρει τις σχέσεις μεταξύ των στοιχείων. Παραδείγματος χάριν, τα στοιχεία μπορούν να αποκαλύψουν ότι οι πελάτες που αγοράζουν τα επιδόρπια χαμηλής περιεκτικότητας σε λιπαρά είναι επίσης μεγάλοι αγοραστές των βοτανικών ενισχύσεων υγείας και ομορφιάς, ή ότι οι καταναλωτές του κρασιού απολαμβάνουν το θέατρο. Ένας αναλυτής στο Wal-Mart, σημείωσε έναν

συσχετισμό μεταξύ των πωλήσεων πανών και των πωλήσεων μπύρας, ο οποίος ήταν ιδιαίτερα ισχυρός τις Παρασκευές. Σε περαιτέρω έρευνα διαπίστωσε ότι οι πατέρες αγόραζαν τις πάνες και έπαιρναν και μια συσκευασία έξι μπυρών συγχρόνως. Η επιχείρηση αποκρίθηκε σε αυτές τις πληροφορίες με το να τοποθετήσει αυτά τα προϊόντα πιο κοντά το ένα στο άλλο. Οι πωλήσεις και των δύο αυξήθηκαν έντονα.

Εξόρυξη δεδομένων στη Marks&Spencer

Η εξόρυξη δεδομένων έχει αποδειχθεί μια επιτυχής στρατηγική για τη βρετανική Marks & Spencer (M&S) Η επιχείρηση παράγει μεγάλους όγκους στοιχείων από δέκα εκατομμύρια πελάτες που εξυπηρετεί κάθε εβδομάδα σε πάνω από 300 καταστήματα. Η οργάνωση θεωρεί ότι η εξόρυξη δεδομένων, αφήνει να χτιστούν σχέσεις με κάθε πελάτη, σε σημείο που όποτε οι μεμονωμένοι πελάτες μπαίνουν σε ένα κατάστημα ο λιανοπωλητής ξέρει ακριβώς ποια προϊόντα πρέπει να προσφέρει προκειμένου να χτιστεί η αποδοτικότητα. Η Marks & Spencer θεωρεί ότι δύο παράγοντες είναι σημαντικοί στην εξόρυξη δεδομένων. Πρώτα είναι η ποιότητα των στοιχείων. Αυτή είναι υψηλότερη όταν είναι γνωστή η ταυτότητα των πελατών, συνήθως ως αποτέλεσμα του ηλεκτρονικού εμπορίου ή της ιδιότητας μέλους προγράμματος επιβράβευσης. Ο δεύτερος είναι να υπάρχουν σαφείς επιχειρησιακοί στόχοι πριν αρχίσει η εξόρυξη δεδομένων. Παραδείγματος χάριν, η Marks & Spencer χρησιμοποιεί την εξόρυξη δεδομένων για να προσδιορίσει τις ομάδες πελατών "υψηλού πλαισίου", "μέσου πλαισίου" και "χαμηλού πλαισίου". Η επιχείρηση έπειτα δημιουργεί προφίλ για τους πελάτες του "υψηλού πλαισίου". Αυτό χρησιμοποιείται για να καθοδηγήσει τις δραστηριότητες διατήρησης πελατών με την κατάλληλα στοχευμένη διαφήμιση και προώθηση. Αυτή η τεχνική μπορεί επίσης να χρησιμοποιηθεί για να κατηγοριοποιηθούν οι πελάτες "μέσου" και "χαμηλού πλαισίου" που έχουν την προοπτική να αναπτυχθούν σε πελάτες ¨"υψηλού πλαισίου"

Τα διαδοχικά σχέδια προκύπτουν συχνά από την εξόρυξη δεδομένων. Οι άνθρωποι που κάνουν εξόρυξη δεδομένων κοιτάζουν για τους «εάν...κατόπιν» κανόνες στη συμπεριφορά πελατών. Παραδείγματος χάριν, μπορεί να βρουν έναν κανόνα όπως: «εάν ένας πελάτης αγοράσει παπούτσια για περπάτημα το Νοέμβριο, κατόπιν υπάρχει μια πιθανότητα 40% ότι θα αγοράσει ένα αδιάβροχο μέσα στους επόμενους έξι μήνες», ή «εάν ένας πελάτης καλέσει το κέντρο πληροφοριών για να ζητήσει πληροφορίες για τα επιτόκια, κατόπιν υπάρχει μια πιθανότητα 50% ότι ο πελάτης θα το σκέφτεται τους επόμενους τρεις μήνες». Κανόνες όπως αυτοί επιτρέπουν στους χρήστες CRM να εφαρμόσουν έγκαιρα μια τακτική.

Στην πρώτη περίπτωση, υπάρχει μια ευκαιρία για πώληση. Αφετέρου, μπορεί να υπάρξει μια ευκαιρία να διατηρηθεί ο πελάτης.

Η εξόρυξη δεδομένων επίσης χρησιμοποιεί την ταξινόμηση. Οι πελάτες μπορούν να είναι σε αμοιβαία αποκλειστικές ομάδες. Παραδείγματος χάριν, να είστε σε θέση να διασταυρώνετε τους υπάρχοντες πελάτες σας στις ομάδες σύμφωνα με την αξία που παράγουν για την επιχείρησή σας. Μπορείτε έπειτα να δημιουργήσετε προφίλ για κάθε ομάδα. Όταν προσδιορίζετε έναν πιθανό νέο πελάτη μπορείτε να κρίνετε σε ποια ομάδα ταιριάζει σύμφωνα με την προοπτική του. Αυτό θα σας δώσει μια ιδέα της πιθανής αξίας της προοπτικής.

Θα μπορούσατε επίσης να ταξινομήσετε τους πελάτες σε πεντάδες ή δεκάδες από την άποψη της σημαντικότητας των πληροφοριών των συναλλαγών όπως την συχνότητα και την χρηματική αξία των αγορών που έχουν κάνει. Αυτό λέγεται ανάλυση RFM. Κατόπιν μπορείτε να πειραματιστείτε με διαφορετικές επεξεργασίες, με την παραγωγή διαφορετικών προσφορών και επικοινωνία με διαφορετικούς τρόπους με τα επιλεγμένα κύτταρα της μήτρας RFM. Μπορεί να βρείτε ότι οι πελάτες που έχουν αγοράσει πιο πρόσφατα, πιο συχνά ή ξοδεύουν περισσότερα ανταποκρίνονται γενικά πιο άμεσα.

Μια άλλη προσέγγιση στην εξόρυξη δεδομένων είναι η ομαδοποίηση. Οι επαγγελματίες του CRM προσπαθούν να συγκεντρώσουν τους πελάτες σε ομάδες. Ο γενικός στόχος της ομαδοποίησης είναι να ελαχιστοποιήσει τις διαφορές μεταξύ των μελών μιας ομάδας όπως επίσης και να μεγιστοποιήσει τις διαφορές μεταξύ των ομάδων. Οι τεχνικές ομαδοποίησης λειτουργούν με τη χρησιμοποίηση μιας συγκεκριμένης σειράς μεταβλητών που εκτελούν τη διαδικασία της ομαδοποίησης. Παραδείγματος χάριν, μπορεί να χρησιμοποιηθούν όλα τα διαθέσιμα στοιχεία συναλλαγής για να παραχθούν τα τμήματα πελατών. Υπάρχουν διάφορες τεχνικές, όπως η ανάλυση συστάδων, η οποία βρίσκει τις κρυμμένες ομάδες. Μόλις διαμορφωθούν οι στατιστικές ομάδες πρέπει να ερμηνευθούν. Οι τομείς της lifestyle αγοράς είναι αποτέλεσμα της ανάλυσης ομάδων σε μεγάλα σύνολα στοιχείων. Οι ετικέτες των ομάδων όπως οι «νέες οικογένειες εργατικής τάξης» ή «τα πλούσια προάστια» χρησιμοποιούνται συχνά για να συλλάβουν την ουσία της ομάδας. Τέλος, η εξόρυξη δεδομένων μπορεί να συμβάλει στο CRM με την παραγωγή προβλέψεων. Οι επαγγελματίες του CRM χρησιμοποιούν το ιστορικό αγοραστικής συμπεριφοράς για να προβλέψουν τη μελλοντική αγοραστική συμπεριφορά και την αξία διάρκειας των πελατών.

Αυτές οι πέντε σημαντικές προσεγγίσεις στην εξόρυξη δεδομένων μπορούν να χρησιμοποιηθούν σε διάφορες ακολουθίες. Παραδείγματος χάριν, θα μπορούσατε να χρησιμοποιήσετε τη συγκέντρωση για να δημιουργήσετε τμήματα πελατών, έπειτα μέσα στα

τμήματα μπορούν να χρησιμοποιηθούν τα στοιχεία συναλλαγών για να προβλεφθούν μελλοντικές αγορές και η αξία διάρκειας των πελατών.

Σύμφωνα με την Gartner Inc., πρωτοπόρα στην αγορά SAS και SPSS, προσφέρει πολλές λύσεις για εξόρυξη δεδομένων που ικανοποιούν τις περισσότερες ανάγκες της αγοράς. Υπάρχουν πολλοί άλλοι προμηθευτές. Οι πιο επιτυχείς προμηθευτές του αναλυτικού CRM προσφέρουν τα ακόλουθα:

- εφαρμογές που υποστηρίζουν τις κοινές αποφάσεις CRM όπως οι σταυροειδείς πωλήσεις (cross-selling) και προβλέψεις αποχώρησης πελατών (customer churn prediction)
- ένα interface κατάλληλο για τους επιχειρησιακούς χρήστες
- η ικανότητα να προσεγγιστούν τα στοιχεία από τις διάφορες πηγές συμπεριλαμβανομένων των στοιχείων από τις αποθήκες δεδομένων , τα datamarts, τα κέντρα κλήσης, το ηλεκτρονικό εμπόριο ή web-tracking συστήματα, καθώς επίσης και τις πηγές στοιχείων τρίτων
- γερά στατιστικά εργαλεία εξόρυξης δεδομένων όπως η ανάλυση συστάδων, δέντρα αποφάσεων και νευρωνικά δίκτυα που μπορούν να παρέχουν αξιόπιστες ιδέες για διαφορετικούς τύπους και όγκους στοιχείων
- εργαλεία αναφοράς (reporting tools) που καθιστούν τα αποτελέσματα της ανάλυσης διαθέσιμα στους ιθύνοντες όπως οι υπεύθυνοι εκστρατειών και οι υπεύθυνοι του τηλεφωνικού κέντρου. (Francis Buttle, 2009)

3.2 Εφαρμογές της εξόρυξης δεδομένων στο CRM

Η διαχείριση των σχέσεων με τους πελάτες εστιάζει φυσικά σε αυτούς που είναι ήδη πελάτες. Ευτυχώς, οι πελάτες είναι η πλουσιότερη πηγή στοιχείων για εξόρυξη. Το καλύτερο επίσης είναι ότι τα στοιχεία που παράγονται από τους πελάτες της επιχείρησης, απεικονίζουν την πραγματική συμπεριφορά του καθενός. Ο πελάτης πληρώνει τους λογαριασμούς του εγκαίρως; Επιταγή ή πιστωτική κάρτα; Πότε ήταν η τελευταία του αγορά; Ποιο προϊόν αγόρασε; Πόσο κόστισε; Πόσες φορές έχει καλέσει ο πελάτης την εξυπηρέτηση πελατών; Πόσες φορές έχουμε καλέσει τον πελάτη; Ποια μέθοδο αποστολής επιλέγει ο πελάτης συχνότερα; Πόσες φορές έχει επιστρέψει μια αγορά; Αυτό το είδος στοιχείων συμπεριφοράς μπορεί να χρησιμοποιηθεί για να αξιολογήσει την πιθανή αξία των πελατών, να αξιολογήσει

τον κίνδυνο αποχώρησης, να αξιολογήσει τον κίνδυνο ότι θα σταματήσουν να πληρώνουν τους λογαριασμούς τους, και για το αν θα προλάβουν τις μελλοντικές ανάγκες τους.

Σύμφωνα με τους Berry & Linoff (2004) μια κοινή εφαρμογή της εξόρυξης δεδομένων είναι η διαμόρφωση ανταπόκρισης. Οι πληροφορίες μπορούν να χρησιμοποιηθούν για να βελτιώσουν το ποσοστό ανταπόκρισης μιας εκστρατείας, αλλά δεν είναι, από μόνες τους, αρκετές να καθορίσουν την αποδοτικότητα της εκστρατείας. Ο υπολογισμός της αποδοτικότητας της εκστρατείας απαιτεί την εμπιστοσύνη στις εκτιμήσεις του ελλοχεύοντος ποσοστού ανταπόκρισης σε μια μελλοντική εκστρατεία, τις εκτιμήσεις των μέσων μεγεθών παραγγελίας που συνδέονται με την ανταπόκριση, και τους προϋπολογισμούς δαπανών για την εκπλήρωση και για την ίδια την εκστρατεία. Μια πελατοκεντρική χρήση των αποτελεσμάτων ανταπόκρισης είναι να επιλεχτεί η καλύτερη εκστρατεία για κάθε πελάτη μεταξύ διάφορων ανταγωνιστικών εκστρατειών. Αυτή η προσέγγιση αποφεύγει το συνηθισμένο πρόβλημα των ανεξάρτητων, -βασισμένων στα αποτελέσματα- εκστρατειών, οι οποίες τείνουν να επιλέξουν τους ίδιους ανθρώπους κάθε φορά.

Είναι σημαντικό κάποιος να διακρίνει μεταξύ της δυνατότητας ενός προτύπου να αναγνωριστούν οι άνθρωποι που ενδιαφέρονται για ένα προϊόν ή μια υπηρεσία και τη δυνατότητά του να αναγνωρίσει τους ανθρώπους που κινούνται για να κάνουν μια αγορά βασισμένοι σε μια ιδιαίτερη εκστρατεία ή μια προσφορά. Η διαφορική ανάλυση ανταπόκρισης προσφέρει έναν τρόπο να προσδιοριστούν οι τομείς αγοράς όπου μια εκστρατεία θα ασκήσει τη μέγιστη επίδραση. Τα διαφορικά πρότυπα ανταπόκρισης επιδιώκουν να μεγιστοποιήσουν τη διαφορά στην ανταπόκριση μεταξύ μιας ομάδας παρατήρησης και μιας ομάδας ελέγχου παρά την προσπάθεια να μεγιστοποιηθεί η ίδια η ανταπόκριση. Οι πληροφορίες για τους τρέχοντες πελάτες μπορούν να χρησιμοποιηθούν για να προσδιορίσουν τις πιθανές προοπτικές με την εύρεση των προβλεπτών των επιθυμητών εκβάσεων στις πληροφορίες που ήταν γνωστές για τους τρέχοντες πελάτες προτού να γίνουν πελάτες. Αυτό το είδος της ανάλυσης είναι πολύτιμο για την επιλογή των καναλιών απόκτησης και των στρατηγικών επαφών καθώς επίσης και για κατάλογους διαλογής. Οι επιχειρήσεις μπορούν να αυξήσουν την αξία των στοιχείων των πελατών τους, αρχίζοντας να ακολουθούν τους πελάτες από την πρώτη αντίδραση τους, ακόμη και προτού να γίνουν πελάτες, και τη συγκέντρωση και την αποθήκευση των πρόσθετων πληροφοριών όταν αποκτιούνται.

Μόλις αποκτηθούν οι πελάτες, η εστίαση μετατοπίζεται στη διαχείριση των σχέσεων με τους πελάτες. Τα διαθέσιμα στοιχεία για τους ενεργούς πελάτες είναι πιο πλούσια από τα διαθέσιμα για τους πιθανούς πελάτες και, επειδή είναι συμπεριφορικής φύσης παρά απλά

γεωγραφικά και δημογραφικά, είναι πιο προφητικά. Η εξόρυξη δεδομένων χρησιμοποιείται για να προσδιορίσει τα πρόσθετα προϊόντα και τις υπηρεσίες που πρέπει να προσφερθούν στους πελάτες με βάση τα τρέχοντα σχέδια χρήσης τους. Μπορεί επίσης να προτείνει τον καλύτερο χρόνο να γίνουν διασταυρωμένες ή αναβαθμισμένες προσφορές.

Ένας από τους στόχους ενός διοικητικού προγράμματος σχέσεων με τους πελάτες είναι να διατηρηθούν οι πολύτιμοι πελάτες. Η εξόρυξη δεδομένων μπορεί να βοηθήσει στο προσδιορισμό των πελατών που είναι πολυτιμότεροι και αξιολογούν τον κίνδυνο εθελοντικής ή ακούσιας αποχώρησης που συνδέεται με κάθε πελάτη. Οπλισμένες με αυτές τις πληροφορίες, οι επιχειρήσεις μπορούν να στοχεύσουν στις προσφορές διατήρησης στους πελάτες που είναι και πολύτιμοι και σε κίνδυνο, και λαμβάνουν μέτρα για να προστατευθούν από τους πελάτες που είναι πιθανό να μην είναι φερέγγυοι.

Από την προοπτική της εξόρυξης δεδομένων, η διαμόρφωση αποχώρησης μπορεί να προσεγγιστεί ως είτε πρόβλημα πρόβλεψης είτε μέσω της ανάλυσης επιβίωσης. Υπάρχουν πλεονεκτήματα και μειονεκτήματα και στις δύο προσεγγίσεις. Η δυαδική προσέγγιση έκβασης λειτουργεί καλά βραχυχρόνια, ενώ η προσέγγιση ανάλυσης επιβίωσης μπορεί να χρησιμοποιηθεί για να καταστήσει τις προβλέψεις μακροχρόνια και παρέχει τη διορατικότητα στην πίστη και την αξία των πελατών.

Ταιριάζοντας τις εκστρατείες με τους πελάτες

Τα ίδια αποτελέσματα ανταπόκρισης που χρησιμοποιούνται για να βελτιστοποιήσουν τον προϋπολογισμό για μια αποστολή σε πιθανούς πελάτες είναι ακόμα πιο χρήσιμα με τους υπάρχοντες πελάτες όπου μπορούν να χρησιμοποιηθούν για να προσαρμόσουν το μίγμα μάρκετινγκ των μηνυμάτων που μια επιχείρηση κατευθύνει στους υπάρχοντες πελάτες της. Το μάρκετινγκ δεν σταματά μόλις αποκτηθούν οι πελάτες. Υπάρχουν εκστρατείες συμπληρωματικών πωλήσεων (cross-sell campaigns), εκστρατείες αναβαθμιζόμενων πωλήσεων (Up-sell campaigns), εκστρατείες υποκίνησης χρήσης, προγράμματα επιβράβευσης κλπ. Αυτές οι εκστρατείες μπορούν να θεωρηθούν ανταγωνιστικές για την πρόσβαση στους πελάτες.

Όταν κάθε εκστρατεία εξετάζεται μεμονωμένα, και σε όλους τους πελάτες δίνονται ποσοστά ανταπόκρισης, αυτό που συμβαίνει συνήθως είναι ότι μια παρόμοια ομάδα πελατών παίρνει υψηλά αποτελέσματα για πολλές από τις εκστρατείες. Μερικοί πελάτες ανταποκρίνονται περισσότερο από άλλους, ένα γεγονός που απεικονίζεται στα αποτελέσματα. Αυτή η προσέγγιση οδηγεί στη φτωχή διαχείριση των σχέσεων με τους πελάτες. Η ομάδα που σημείωσε υψηλά αποτελέσματα βομβαρδίζεται με μηνύματα,

ενοχλείται και γίνεται αδιάφορη. Εν τω μεταξύ, άλλοι πελάτες δεν λαμβάνουν νέα από την επιχείρηση ποτέ και έτσι δεν ενθαρρύνονται για να επεκτείνουν τις σχέσεις τους. Μια εναλλακτική λύση είναι να σταλεί ένας περιορισμένος αριθμός μηνυμάτων σε κάθε πελάτη, χρησιμοποιώντας τα αποτελέσματα για να αποφασίσει ποια μηνύματα είναι τα πιο κατάλληλα για τον καθένα.

Κατάτμηση της βάσης πελατών

Η κατάτμηση της βάσης πελατών είναι μια δημοφιλής εφαρμογή της εξόρυξης δεδομένων. Ο σκοπός της κατάτμησης είναι να προσαρμοστούν τα προϊόντα, οι υπηρεσίες, και τα μηνύματα μάρκετινγκ σε κάθε τομέα. Τα τμήματα πελατών έχουν βασιστεί παραδοσιακά στην έρευνα αγοράς και τα δημογραφικά στοιχεία. Μπορεί να υπάρχει ένα "νέο και ενιαίο" τμήμα ή ένα "πιστό τμήμα." Το πρόβλημα με τα τμήματα που είναι βασισμένα στην έρευνα αγοράς είναι ότι είναι δύσκολο να είναι γνωστό πώς να τα εφαρμόσει σε όλους τους πελάτες που δεν ήταν μέρος της έρευνας. Το πρόβλημα με τα τμήματα πελατών βασισμένα στα δημογραφικά στοιχεία είναι ότι όχι όλοι "οι νέοι και ελεύθεροι" ή οι "εργένηδες" πραγματικά έχουν τις ίδιες προτιμήσεις και χαρακτηριστικά με το τμήμα τους. Η προσέγγιση της εξόρυξης δεδομένων είναι να προσδιοριστούν τα τμήματα με τις ίδιες συμπεριφορές.

Συνδυάζοντας τους τομείς της έρευνας αγοράς με τα συμπεριφορικά στοιχεία

Μια από τις μεγάλες προκλήσεις με την παραδοσιακή, βασισμένη στην έρευνα, έρευνα αγοράς είναι ότι παρέχει πολλές πληροφορίες για μερικούς πελάτες. Εντούτοις, για να χρησιμοποιηθούν τα αποτελέσματα της έρευνας αγοράς αποτελεσματικά συχνά απαιτούνται τα χαρακτηριστικά όλων των πελατών. Η έρευνα αγοράς μπορεί να βρει ενδιαφέροντες τομείς πελατών. Αυτοί πρέπει έπειτα να προβληθούν επάνω στην βάση πελατών που χρησιμοποιεί τα διαθέσιμα στοιχεία. Τα συμπεριφορικά στοιχεία μπορούν να είναι ιδιαίτερα χρήσιμα για αυτό. Τέτοια συμπεριφορικά στοιχεία συνοψίζονται από το ιστορικό συναλλαγών και τιμολόγησης. Μια απαίτηση της έρευνας αγοράς είναι ότι οι πελάτες πρέπει να προσδιοριστούν έτσι ώστε η συμπεριφορά των συμμετεχόντων στην έρευνα αγοράς να είναι γνωστή. Οι περισσότερες από τις κατευθυνόμενες τεχνικές εξόρυξης δεδομένων που συζητούνται σε αυτή την εργασία μπορούν να χρησιμοποιηθούν για να χτίσουν ένα πρότυπο ταξινόμησης για να τοποθετήσουν τους ανθρώπους στα τμήματα που είναι βασισμένα στα διαθέσιμα στοιχεία. Αυτό που απαιτείται είναι ένα σύνολο κατάρτισης των πελατών που έχουν ήδη ταξινομηθεί. Το πόσο καλά θα δουλέψει αυτό εξαρτάται κατά ένα μεγάλο μέρος

από το βαθμό στον όποιο τα τμήματα πελατών υποστηρίζονται πραγματικά από τη συμπεριφορά των πελατών.

Προβλέποντας ποιοι δεν θα είναι φερέγγυοι

Το να μάθει μια επιχείρηση να αποφεύγει τους κακούς πελάτες (και να παρατηρεί όταν οι καλοί πελάτες είναι έτοιμοι να γίνουν κακοί) είναι τόσο σημαντικές όσο η διατήρηση των καλών πελατών. Οι περισσότερες επιχειρήσεις των οποίων οι εργασίες τους εκθέτουν σε πιστωτικό κίνδυνο, ελέγχουν την πιστωτική ικανότητα των πελατών ως τμήμα της διαδικασίας απόκτησης, αλλά το μοντέλο κινδύνου δεν τελειώνει με την απόκτηση του πελάτη.

Η αξιολόγηση του πιστωτικού κινδύνου στους υπάρχοντες πελάτες είναι ένα πρόβλημα για οποιαδήποτε επιχείρηση παρέχει μια υπηρεσία που οι πελάτες πληρώνουν μετά. Υπάρχει πάντα η πιθανότητα, μερικοί πελάτες που θα λάβουν την υπηρεσία να αποτύχουν έπειτα να πληρώσουν για αυτήν. Οι συνδρομές εφημερίδων, η τηλεφωνικές υπηρεσίες, το αέριο και το ηλεκτρικό, και η υπηρεσία καλωδιακής τηλεόρασης είναι μεταξύ των πολλών υπηρεσιών που πληρώνονται συνήθως αφότου έχουν χρησιμοποιηθεί. Φυσικά, οι πελάτες που αποτυγχάνουν να πληρώσουν για αρκετό καιρό αποκόπτονται.

Έως τότε μπορούν να οφείλουν μεγάλα ποσά χρημάτων που πρέπει να διαγραφούν. Με την έγκαιρη προειδοποίηση από ένα μοντέλο πρόβλεψης, μια επιχείρηση μπορεί να λάβει μέτρα για να προστατευθεί. Αυτά τα βήματα μπορεί να περιλαμβάνουν τον περιορισμό της πρόσβασης στην υπηρεσία ή μειώνοντας το χρονικό διάστημα μεταξύ μιας πληρωμής που καθυστερεί και τη διακοπή της υπηρεσίας. Η ακούσια αποχώρηση, όπως η λήξη των υπηρεσιών για τη μη πληρωμή όπως καλείται μερικές φορές, μπορεί να διαμορφωθεί με πολλούς τρόπους. Συχνά, η ακούσια αποχώρηση θεωρείται ως δυαδική έκβαση σε κάποιο σταθερό χρονικό διάστημα, οπότε σε αυτή την περίπτωση τεχνικές όπως η λογιστική παλινδρόμηση και τα δέντρα απόφασης είναι κατάλληλα.

Μια από τις μεγάλες διαφορές μεταξύ της εκούσιας και της ακούσιας αποχώρησης είναι ότι η ακούσια αποχώρηση περιλαμβάνει συχνά περίπλοκες επιχειρησιακές διαδικασίες, καθώς οι λογαριασμοί περνούν από τα διαφορετικά στάδια ύπαρξης. Κατά τη διάρκεια του χρόνου, οι επιχειρήσεις μπορούν να "πειράξουν" τους κανόνες που καθοδηγούν τις διαδικασίες για να ελέγξουν το χρηματικό ποσό που τους οφείλεται. Κατά την έρευνα των ακριβών αριθμών βραχυπρόθεσμα, η διαμόρφωση του κάθε βήματος στις επιχειρησιακές διαδικασίες μπορεί να είναι η καλύτερη προσέγγιση.

Καθορισμός της αξίας του πελάτη

Οι υπολογισμοί της αξίας των πελατών είναι αρκετά σύνθετοι και ενώ η εξόρυξη δεδομένων διαδραματίζει έναν ρόλο, οι υπολογισμοί της αξίας των πελατών είναι κατά ένα μεγάλο μέρος θέμα των σωστών οικονομικών ορισμών. Μια φαινομενικά απλή δήλωση της αξίας του πελάτη είναι τα συνολικά έσοδα λόγω του πελάτη μείον το συνολικό κόστος διατήρησης του. Αλλά πόσο κέρδος θα έπρεπε να αποδοθεί σε έναν πελάτη; Είναι τι έχει ξοδέψει στο σύνολο μέχρι σήμερα; Τι ξόδεψε αυτό το μήνα; Τι αναμένουμε να ξοδέψει κατά τη διάρκεια του επόμενου έτους; Πώς θα έπρεπε τα έμμεσα έσοδα όπως έσοδα από διαφήμιση και η ενοικίαση καταλόγων να διατεθούν στους πελάτες; Οι δαπάνες είναι ακόμα πιο προβληματικές. Οι επιχειρήσεις έχουν όλα τα είδη των δαπανών που μπορούν να διατεθούν στους πελάτες με τους ιδιαίτερους τρόπους. Ακόμη και αγνοώντας τις διατιθέμενες δαπάνες και εξετάζοντας μόνο τις άμεσες δαπάνες, τα πράγματα μπορούν ακόμα να είναι αρκετά μπερδεμένα. Είναι δίκαιο να κατηγορηθούν οι πελάτες για τις δαπάνες πέρα από τις οποίες που έχουν κάποιο έλεγχο; Δύο πελάτες στο διαδίκτυο παραγγέλνουν τα ίδια ακριβώς εμπορεύματα και στους δύο υπόσχεται δωρεάν παράδοση. Κάποιος που ζει πιο μακριά από την αποθήκη εμπορευμάτων μπορεί να κοστίσει περισσότερα για την αποστολή, αλλά είναι πραγματικά λιγότερο πολύτιμος πελάτης; Οι εταιρίες παροχής κινητής τηλεφωνίας βρίσκονται αντιμέτωπες με ένα παρόμοιο πρόβλημα. Οι περισσότερες διαφημίζουν πια κοινές τιμές διεθνώς. Οι δαπάνες των παροχών δεν είναι ομοιόμορφες όταν δεν είναι κύριοι ολόκληρου του δικτύου. Μερικές από τις κλήσεις ταξιδεύουν πέρα από το δίκτυο της επιχείρησης. Άλλες κλήσεις ταξιδεύουν πέρα από τα δίκτυα των ανταγωνιστών που χρεώνουν υψηλά ποσοστά. Μπορεί η επιχείρηση να αυξήσει την αξία των πελατών με την προσπάθεια να αποθαρρυνθούν οι πελάτες από την επίσκεψη ορισμένων γεωγραφικών περιοχών; Μόλις όλα αυτά τα προβλήματα έχουν τακτοποιηθεί, και μια επιχείρηση έχει συμφωνήσει στον καθορισμό της αναδρομικής αξίας των πελατών, η εξόρυξη δεδομένων μπαίνει στο παιχνίδι προκειμένου να υπολογιστεί η ενδεχόμενη αξία των πελατών. Αυτό φτάνει στον υπολογισμό του εισοδήματος που ένας πελάτης θα φέρει μέσα σε ένα διάστημα και έπειτα υπολογίζει το υπόλοιπο χρόνο ζωής του πελάτη.

Με τους υπάρχοντες πελάτες, ένα σημαντικό κομμάτι στη διαχείριση των σχέσεων με τους πελάτες είναι να αυξήσει την αποδοτικότητα των πελατών μέσω των διασταυρούμενων και αναβαθμισμένων πωλήσεων. Η εξόρυξη δεδομένων χρησιμοποιείται για να καταλάβουμε τι να προσφέρουμε, σε ποιον και πότε να το προσφέρουμε.

Βρίσκοντας το σωστό χρόνο για μια προσφορά

Η Charles Schwab, η επιχείρηση επενδύσεων, ανακάλυψε ότι οι πελάτες συνήθως ανοίγουν ένα λογαριασμό με μερικές χιλιάδες δολάρια ακόμα κι αν έχουν αρκετά περισσότερα σε τράπεζες και σε επενδύσεις. Φυσικά, η Schwab θα επιθυμούσε να προσελκύσει και μερικούς από άλλα ισοζύγια. Με την ανάλυση των ιστορικών στοιχείων, ανακάλυψαν ότι οι πελάτες που μετέφεραν μεγάλα ισοζύγια στους λογαριασμούς επένδυσης το έκαναν συνήθως κατά τη διάρκεια των πρώτων μηνών αφότου άνοιξαν τον πρώτο λογαριασμό τους. Μετά από μερικούς μήνες, υπήρξε μια μικρή επιστροφή στην προσπάθεια να κάνουν τους πελάτες να κινηθούν σε μεγάλα ισοζύγια. Το παράθυρο ήταν κλειστό. Όπως έμαθε από αυτό, η Schwab μετατόπισε τη στρατηγική της από την αποστολή ενός σταθερού ρεύματος παρακλήσεων σε όλο τον κύκλο ζωής των πελατών σε συγκεντρωμένες προσπάθειες κατά τη διάρκεια των πρώτων μηνών. Μια μεγάλη εφημερίδα με συνδρομές στις καθημερινές και κυριακάτικες εκδόσεις της, παρατήρησε ένα παρόμοιο σχέδιο. Εάν ένας συνδρομητής της κυριακάτικης εφημερίδας αναβαθμίσει τη συνδρομή του στην καθημερινή και την κυριακάτικη, συμβαίνει συνήθως νωρίς στη σχέση. Ένας πελάτης που είναι ευχαριστημένος μόνο από την κυριακάτικη εφημερίδα για χρόνια, είναι λιγότερο πιθανό να αλλάξει τις συνήθειές του .

Υποβολή συστάσεων

Μια προσέγγιση στην διασταυρωμένη πώληση χρησιμοποιεί τους κανόνες ένωσης. Χρησιμοποιείται για να βρει τις συστάδες των προϊόντων που πωλούνται συνήθως μαζί ή τείνουν να αγοραστούν από το ίδιο πρόσωπο κατά τη διάρκεια του χρόνου. Οι πελάτες που έχουν αγοράσει μερικά, αλλά όχι όλα τα μέλη μιας συστάδας έχουν καλές προοπτικές για τα υπόλοιπα στοιχεία. Αυτή η προσέγγιση λειτουργεί για τα προϊόντα λιανικής πώλησης όπου υπάρχουν πολλές τέτοιες συστάδες, αλλά είναι λιγότερο αποτελεσματική στις περιοχές όπως οι οικονομικές υπηρεσίες όπου υπάρχουν λιγότερα προϊόντα και πολλοί πελάτες έχουν ένα παρόμοιο μίγμα, και το μίγμα καθορίζεται συχνά από τη συσσώρευση προϊόντων και τις προηγούμενες προσπάθειες μάρκετινγκ. (Berry & Linoff, 2004)

3.3 Ανάλυση καλαθιών αγοράς και κανόνες ένωσης

Για να αναπτυχθούν οι θεμελιώδεις ιδέες της ανάλυσης καλαθιών αγοράς, θα πρέπει πρώτα να εξετάσουμε την εικόνα του καλαθιού αγοράς το οποίο μπορεί να περιέχει διάφορα προϊόντα που αγοράζονται από μια γρήγορη επίσκεψη στην αγορά. Ένα τέτοιο καλάθι μπορεί να περιέχει μια ποικιλία προϊόντων όπως χυμό πορτοκαλιού, μπανάνες, μη αλκοολούχο ποτό, καθαριστικό παραθύρων, και απορρυπαντικό. Ένα καλάθι μας λέει τι αγόρασε συγχρόνως ένας πελάτης. Επομένως ένας πλήρης κατάλογος αγορών που γίνονται από όλους τους πελάτες παρέχει πολύ περισσότερες πληροφορίες, επίσης περιγράφει το σημαντικότερο μέρος των πωλήσεων μιας λιανικής επιχείρησης δηλαδή τι προϊόντα αγοράζουν οι πελάτες και πότε.

Κάθε πελάτης αγοράζει ένα διαφορετικό σύνολο προϊόντων, σε διαφορετικές ποσότητες, και σε διαφορετικούς χρόνους. Κατ' επέκταση η ανάλυση καλαθιών αγοράς χρησιμοποιεί τις πληροφορίες που έχουν σχέση με το τι αγόρασε ο κάθε πελάτης για να παρέχει τη διορατικότητα για το ποιοι είναι και γιατί κάνουν ορισμένες αγορές. Επίσης η ανάλυση καλαθιών αγοράς παρέχει διορατικότητα στα εμπόρευμα, λέγοντάς μας ποια προϊόντα τείνουν να αγοραστούν μαζί και ποια επιδέχονται καλύτερα την προώθηση

Η τεχνική εξόρυξης δεδομένων που συνδέεται όλο και περισσότερο με την ανάλυση καλαθιών αγοράς είναι η αυτόματη παραγωγή των κανόνων ένωσης. Στην πραγματικότητα οι κανόνες ένωσης αντιπροσωπεύουν σχέδια στα στοιχεία χωρίς έναν διευκρινισμένο στόχο. Επομένως υπό αυτήν τη μορφή, είναι ένα παράδειγμα της μη κατευθυνόμενης εξόρυξης δεδομένων. Τέλος αφήνεται στην ανθρώπινη ερμηνεία να αποφασίσει για το αν τα σχέδια έχουν νόημα

Οι κανόνες ένωσης προήλθαν αρχικά από στοιχεία των σημείων πώλησης που περιγράφουν ποια προϊόντα αγοράζονται από κοινού. Αν και οι ρίζες τους βρίσκονται στην ανάλυση συναλλαγών των σημείων πώλησης, οι κανόνες ένωσης μπορούν να εφαρμοστούν και έξω από τη λιανική βιομηχανία για να βρουν τις σχέσεις μεταξύ άλλων τύπων "καλαθιών." Μερικά παραδείγματα πιθανών εφαρμογών είναι:

- Τα προϊόντα που αγοράζονται με πιστωτική κάρτα, όπως τα ενοικιαζόμενα αυτοκίνητα και τα δωμάτια ξενοδοχείων, παρέχουν διορατικότητα για το επόμενο προϊόν που οι πελάτες είναι πιθανό να αγοράσουν.

- Οι προαιρετικές υπηρεσίες που αγοράζονται από τους πελάτες τηλεπικοινωνιών (αναμονή κλήσης, ταχεία κλήση, και τα λοιπά) βοηθούν στο να καθοριστεί πώς να συσσωρευτούν αυτές τις υπηρεσίες για να μεγιστοποιηθούν τα έσοδα.
- Οι τραπεζικές υπηρεσίες που χρησιμοποιούνται από τους λιανικούς πελάτες (απολογισμοί αγοράς χρημάτων, υπηρεσίες επένδυσης, δάνεια αυτοκινήτων κλπ.) προσδιορίζουν πελάτες που πιθανώς να θελήσουν άλλες υπηρεσίες.
- Οι ασυνήθιστοι συνδυασμοί ασφαλιστικών αξιώσεων μπορούν να είναι ένα σημάδι απάτης και μπορούν να προκαλέσουν περαιτέρω έρευνα.
- Τα ιατρικά ιστορικά των ασθενών μπορούν να δώσουν ενδείξεις των πιθανών περιπλοκών βασισμένα σε ορισμένους συνδυασμούς θεραπειών.

Οι κανόνες ένωσης συχνά αποτυγχάνουν να εκπληρώσουν τις προσδοκίες, παραδείγματος χάριν, δεν είναι καλή επιλογή για την οικοδόμηση ενός προτύπου σταυροειδούς πώλησης στις βιομηχανίες όπως οι λιανικές τραπεζικές εργασίες, επειδή οι κανόνες συνήθως καταλήγουν σε προηγούμενες προωθήσεις μάρκετινγκ. Επίσης, στις λιανικές τραπεζικές εργασίες, οι πελάτες αρχίζουν με έναν απλό λογαριασμό και έπειτα αποκτούν έναν λογαριασμό αποταμίευσης. Αυτή η διαφοροποίηση μεταξύ των προϊόντων δεν εμφανίζεται έως ότου οι πελάτες έχουν περισσότερα προϊόντα. (Berry and Linoff, 2004)

3.3.1 Καθορισμός της ανάλυσης καλαθιών αγοράς

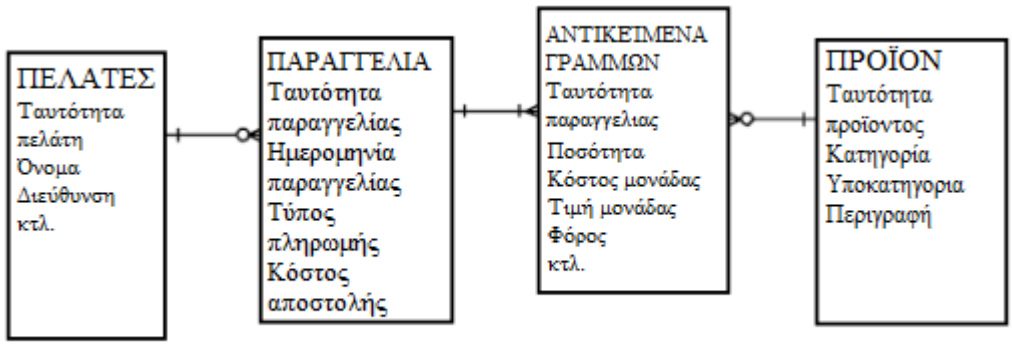
Η ανάλυση καλαθιών αγοράς δεν αναφέρεται σε μια ενιαία τεχνική, αναφέρεται σε ένα σύνολο επιχειρησιακών προβλημάτων σχετικών με την κατανόηση των στοιχείων συναλλαγής στα σημεία πώλησης. Η πιο κοινή τεχνική είναι οι κανόνες ένωσης. Όμως πριν τους κανόνες ένωσης θα αναφερθούμε στα δεδομένα των καλαθιών αγοράς.

Τα τρία επίπεδα των δεδομένων των καλαθιών αγοράς

Τα στοιχεία καλαθιών αγοράς είναι στοιχεία συναλλαγής που περιγράφουν τρεις πλήρως διαφορετικές οντότητες:

- **Πελάτες**
- **Παραγγελίες** (επίσης αποκαλούμενες *αγορές* ή *σύνολα καλαθιών.)*
- **Αντικείμενα ή προϊόντα**

Σε μια σχεσιακή βάση δεδομένων, η δομή δεδομένων για τα στοιχεία καλαθιών αγοράς φαίνεται συχνά παρόμοια με το σχήμα 3.1. Αυτή η δομή δεδομένων περιλαμβάνει τέσσερις σημαντικές οντότητες.



Σχήμα 3.1 Ένα μοντέλο δεδομένων για τα δεδομένα καλαθιών αγοράς σε επίπεδο συναλλαγής έχει συνήθως τρεις πίνακες έναν για τον πελάτη, έναν για την παραγγελία, και έναν για την γραμμή παραγγελίας.(Berry and Linoff, 2004)

Η παραγγελία είναι η θεμελιώδης δομή δεδομένων για τα αρχεία των καλαθιών αγοράς. Συγκεκριμένα μια παραγγελία αντιπροσωπεύει ένα ενιαίο γεγονός αγορών από έναν πελάτη. Ως εκ τούτου αυτό αντιστοιχεί σε έναν πελάτη που παραγγέλνει διάφορα προϊόντα σε έναν ιστοχώρο ή σε έναν πελάτη που αγοράζει ένα καλάθι από το παντοπωλείο ή σε μια αγορά διάφορων αντικειμένων από έναν κατάλογο. Επίσης αυτό περιλαμβάνει το συνολικό ποσό της αγοράς, πρόσθετες δαπάνες, τον τύπο πληρωμής, και οποιοσδήποτε άλλο στοιχείο είναι σχετικό με τη συναλλαγή.

Τα μεμονωμένα προϊόντα της παραγγελίας αντιπροσωπεύονται χωριστά ως αντικείμενα γραμμών. Κατ' επέκταση αυτά τα δεδομένα περιλαμβάνουν την τιμή που καταβλήθηκε για την αγορά του αντικειμένου, τον αριθμό των αντικειμένων, επίσης εάν ο φόρος πρέπει να χρεωθεί, και ίσως το κόστος. Επίσης ο πίνακας αντικειμένων έχει μια σύνδεση με έναν πίνακα αναφοράς προϊόντων, ο οποίος παρέχει περισσότερες περιγραφικές πληροφορίες για κάθε προϊόν. Αυτές οι περιγραφικές πληροφορίες πρέπει να περιλαμβάνουν την ιεραρχία των προϊόντων και άλλες πληροφορίες που μπορούν να αποδειχθούν πολύτιμες για την ανάλυση.

Ο πίνακας πελατών είναι ένας προαιρετικός πίνακας και πρέπει να είναι διαθέσιμος όταν μπορεί ένας πελάτης να προσδιοριστεί, παραδείγματος χάριν, σε έναν ιστοχώρο που απαιτείτε η εγγραφή ή όταν χρησιμοποιεί ο πελάτης μια πιστωτική κάρτα συγγένειας (affinity card) κατά τη διάρκεια της συναλλαγής. Αν και ο πίνακας πελατών μπορεί να έχει πολλούς ενδιαφέροντες τομείς, το ισχυρότερο στοιχείο είναι η ίδια η ταυτότητα, επειδή η ταυτότητα μπορεί να συνδέει τις συναλλαγές με το πέρασμα του χρόνου.

Και τα τρία επίπεδα δεδομένων καλαθιών αγοράς είναι σημαντικά. Παραδείγματος χάριν, για να γίνουν κατανοητές οι παραγγελίες, υπάρχουν μερικά βασικά μέτρα:

- Ποιος είναι ο μέσος αριθμός παραγγελιών ανά πελάτη;
- Ποιος είναι ο μέσος αριθμός μοναδικών προϊόντων ανά παραγγελία;
- Ποιος είναι ο μέσος αριθμός προϊόντων ανά παραγγελία;
- Για ένα δεδομένο προϊόν, ποιο είναι το ποσοστό των πελατών που έχουν αγοράσει το προϊόν;
- Για ένα δεδομένο προϊόν, ποιος είναι ο μέσος αριθμός παραγγελιών ανά πελάτη που περιλαμβάνουν το προϊόν;
- Για ένα δεδομένο προϊόν, ποια είναι η μέση ποσότητα που αγοράζεται σε μια παραγγελία;

Αυτά τα μέτρα δίνουν μία ευρεία διορατικότητα στην επιχείρηση. Επίσης σε μερικές περιπτώσεις, υπάρχει ένας μικρός αριθμός επαναλαμβανόμενων πελατών, αυτό δείχνει μια επιχειρησιακή ευκαιρία δηλαδή να αυξηθεί ο αριθμός πωλήσεων ανά πελάτη. Επίσης, ο αριθμός προϊόντων ανά παραγγελία μπορεί να είναι επαναλαμβανόμενος, προτείνοντας έτσι μια ευκαιρία για μία διαγώνιο-πώληση κατά τη διάρκεια της διαδικασίας μιας παραγγελίας. Επίσης μπορεί να είναι χρήσιμο να συγκριθούν αυτά τα μέτρα μεταξύ τους. Εξάλλου έχει διαπιστωθεί ότι ο αριθμός παραγγελιών είναι συχνά ένας χρήσιμος τρόπος διαφοροποίησης μεταξύ των πελατών οι καλοί πελάτες παραγγέλνουν σαφώς συχνότερα από ότι οι όχι τόσο καλοί πελάτες.

Χαρακτηριστικά παραγγελίας

Οι αγορές πελατών έχουν πρόσθετα ενδιαφέροντα χαρακτηριστικά. Παραδείγματος χάριν, το μέσο μέγεθος παραγγελίας ποικίλλει ανάλογα του χρόνου και της περιοχής και είναι χρήσιμη η παρακολούθηση αυτών για να γίνονται αντιληπτές οι αλλαγές στο επιχειρησιακό περιβάλλον.

Κάποιες πληροφορίες, εν τούτοις, είναι απαραίτητο να αντληθούν από τα στοιχεία συναλλαγών. Εάν για παράδειγμα διασπάσουμε τις συναλλαγές ανάλογα με το μέγεθος της παραγγελίας και την πιστωτική κάρτα που χρησιμοποιήθηκε για την πληρωμή (π.χ. Visa, Master Card, American Express) και γίνει άντληση πληροφοριών τότε θα διαπιστώσουμε ότι όσο μεγαλύτερη η παραγγελία, τόσο μεγαλύτερο είναι το μέσο ποσό αγορών, ανεξάρτητα

από την πιστωτική κάρτα που χρησιμοποιείται. Επίσης, η χρήση ενός τύπου πιστωτικών καρτών (American express), συνδέεται συχνότερα με μεγαλύτερες παραγγελίες.

Για τις αγορές μέσο διαδικτύου και τις συναλλαγές ταχυδρομικών παραγγελιών, οι πρόσθετες πληροφορίες μπορούν επίσης να συγκεντρωθούν στο σημείο της πώλησης:

- Η παραγγελία έγινε με περιτύλιγμα δώρου;
- Η παραγγελία πηγαίνει στην ίδια διεύθυνση με τη διεύθυνση τιμολόγησης;
- Ο αγοραστής δέχτηκε ή όχι μια προσφορά διαγώνιας πώλησης;

Φυσικά, η συγκέντρωση των πληροφοριών στο σημείο της πώλησης και η διαθεσιμότητα για την ανάλυση είναι δύο διαφορετικά πράγματα. Εντούτοις, το αν χρησιμοποιείτε περιτύλιγμα δώρων και η ανταπόκριση σε μια σταυροειδή-πώληση είναι δύο πολύ χρήσιμα πράγματα που μπορούμε να γνωρίζουμε για τους πελάτες. Επίσης η εύρεση σχεδίων μέσο αυτών των πληροφοριών απαιτεί εξαρχής την συλλογή των πληροφοριών και έπειτα την μεταφορά τους προς ένα περιβάλλον εξόρυξης δεδομένων. (Berry and Linoff. 2004)

Δημοτικότητα προϊόντων

Ποια είναι τα δημοφιλέστερα προϊόντα; Εντούτοις η γνώση των πωλήσεων ενός μεμονωμένου προϊόντος είναι μόνο η αρχή. Υπάρχουν και άλλες σχετικές ερωτήσεις:

- Ποιο είναι το προϊόν που συναντάται συχνότερα σε μια παραγγελία ενός προϊόντος;
- Ποιο είναι προϊόν που συναντάται συχνότερα σε μια παραγγελία πολλών προϊόντων;
- Ποιο είναι προϊόν που συναντάται συχνότερα σε αγορές των πελατών που παραγγέλνουν επανειλημμένα;
- Πώς η δημοτικότητα συγκεκριμένων προϊόντων έχει αλλάξει κατά τη διάρκεια του χρόνου;
- Πώς η δημοτικότητα ενός προϊόντος ποικίλλει ανά περιφέρεια;

Οι πρώτες τρεις ερωτήσεις είναι ιδιαίτερα ενδιαφέρουσες επειδή μπορούν να προτείνουν ιδέες για την ανάπτυξη των σχέσεων πελατών. Επίσης οι κανόνες ένωσης μπορούν να δώσουν απαντήσεις σε αυτές τις ερωτήσεις, ιδιαίτερα όταν χρησιμοποιούνται με εικονικά προϊόντα για να αντιπροσωπεύσουν το μέγεθος της παραγγελίας ή τον αριθμό παραγγελιών που έχει κάνει ένας πελάτης.

Οι τελευταίες δύο ερωτήσεις έχουν να κάνουν με τις διαστάσεις του χρόνου και της γεωγραφίας, οι οποίες είναι πολύ σημαντικές για τις εφαρμογές της ανάλυσης καλαθιών

αγοράς. Συγκεκριμένα τα διαφορετικά προϊόντα έχουν και διαφορετικές συγγένειες σε διαφορετική περιοχή.

Εξετάζοντας τις επεμβάσεις μάρκετινγκ

Η εξέταση μεμονωμένων προϊόντων κατά τη διάρκεια του χρόνου μπορεί να παρέχει μια καλή κατανόηση του τι συμβαίνει με το προϊόν. Όπως για παράδειγμα, των επεμβάσεων του μάρκετινγκ μαζί με τις πωλήσεις προϊόντων κατά τη διάρκεια του χρόνου.

Τα δεδομένα των καλαθιών αγοράς μπορούν να βοηθήσουν στην εξέταση του όγκου των πωλήσεων μετά από μια επέμβαση μάρκετινγκ, όπως επίσης στην εξέταση του αριθμού καλαθιών που περιέχουν το προϊόν. Ομοίως, είναι δυνατόν να εξεταστεί εάν το μέσο μέγεθος των παραγγελιών αυξήθηκε ή μειώθηκε μετά από μια επέμβαση.

Ομαδοποίηση των προϊόντων βάση της χρήσης

Οι ομάδες προϊόντων που εμφανίζονται συχνά μαζί. μπορούν να δώσουν πληροφορίες οι οποίες μπορούν να φανούν πολύ χρήσιμες για την υποβολή συστάσεων πελατών ,όπως για παράδειγμα πελάτες οι οποίοι έχουν αγοράσει κάποια από τα προϊόντα της ομάδας και μπορεί να ενδιαφέρονται και για τα υπόλοιπα. Στο μεμονωμένο επίπεδο προϊόντων, οι κανόνες ένωσης δίνουν μερικές απαντήσεις σε αυτήν την περιοχή. Ειδικότερα, αυτή η τεχνική εξόρυξης δεδομένων καθορίζει ποιο προϊόν ή προϊόντα σε μια αγορά προτείνουν την αγορά άλλων ιδιαίτερων προϊόντων συγχρόνως.

Συνήθως υπάρχουν πολλές διαθέσιμες πληροφορίες σχετικά με τα προϊόντα. Εκτός από την ιεραρχία του προϊόντος, οι πληροφορίες αυτές μπορούν να περιλαμβάνουν το χρώμα των ρούχων, εάν τα τρόφιμα είναι χαμηλών θερμίδων, εάν μια αφίσα περιλαμβάνει ένα κάδρο, και ούτω καθεξής. Αυτές οι περιγραφές παρέχουν έναν πλούτο πληροφοριών, και μπορούν να οδηγήσουν σε χρήσιμα ερωτήματα:

- Τα προϊόντα διαίτης έχουν την τάση να πωλούνται από κοινού;
- Οι πελάτες αγοράζουν ρούχα με παρόμοια χρώματα την ίδια στιγμή;
- Οι πελάτες που αγοράζουν πλαισιωμένες αφίσες προβαίνουν και σε άλλες αγορές;

Εντούτοις με το να είμαστε σε θέση να απαντήσουμε σε αυτά τα ερωτήματα είναι συχνά πιο χρήσιμο από το να γίνει προσπάθεια ομαδοποίησης των προϊόντων, δεδομένου ότι τέτοια απευθείας ερωτήματα συχνά οδηγούν άμεσα σε ενέργειες μάρκετινγκ. (Edelstein H., 2000)

3.3.2 Κανόνες ένωσης

Μια έφεση των κανόνων ένωσης είναι η σαφήνεια και η χρησιμότητα των αποτελεσμάτων, τα οποία είναι υπό μορφή κανόνων που αφορούν τις ομάδες προϊόντων. Πιο συγκεκριμένα υπάρχει μια διαισθητική έφεση σε έναν κανόνα ένωσης επειδή εκφράζει πόσο απτά τα προϊόντα και οι υπηρεσίες ομαδοποιούνται. Ένα παράδειγμα τέτοιου κανόνα είναι, "*εάν ένας πελάτης αγοράσει την υπηρεσία τριπλής κλήσης, τότε εκείνος ο πελάτης θα αγοράσει επίσης και την αναμονή κλήσης,* " ένας τέτοιος κανόνας είναι σαφής. Ακόμα μπορεί να προτείνει ένα συγκεκριμένο σχέδιο δράσης, όπως την συγχώνευση της υπηρεσίας τριπλής κλήσης με την αναμονή κλήσης σε μια ενιαία συσκευασία υπηρεσιών.

Ενώ οι κανόνες ένωσης είναι εύκολο να κατανοηθούν, δεν είναι πάντα χρήσιμοι. Οι ακόλουθοι τρεις κανόνες είναι παραδείγματα πραγματικών κανόνων που παράγονται από πραγματικά στοιχεία:

- Οι πελάτες του Wal-Mart που αγοράζουν κούκλες Barbie έχουν μια πιθανότητα 60 % να αγοράσουν επίσης και ζαχαρωτά.
- Οι πελάτες που κλείνουν συμφωνίες συντήρησης συσκευών είναι πολύ πιθανό να αγοράσουν μεγάλες συσκευές.
- Όταν ένα νέο κατάστημα ανοίγει, ένα από τα προϊόντα που αγοράζεται συχνότερα είναι τα καθαριστικά τουαλετών.(Forbes on September 8, 1997)

Αυτά τα τρία παραδείγματα επεξηγούν τους τρεις κοινούς τύπους κανόνων που παράγονται από τους κανόνες ένωσης: τον *αγώγιμο*, τον *τετριμμένο*, και τ*ον ανεξήγητο*.

Αγώγιμοι κανόνες

Ένας χρήσιμος κανόνας περιέχει και υψηλής ποιότητας, αγώγιμες πληροφορίες. Μόλις βρεθεί το σχέδιο, δεν είναι ιδιαίτερα δύσκολο να δικαιολογηθεί, και η αφήγηση μιας ιστορίας μπορεί να οδηγήσει σε νέες ιδέες και στη δράση. Για παράδειγμα ας φανταστούμε ότι μια οικογένεια πάει για ψώνια. Ο σκοπός είναι η αγορά ενός δώρου για την μικρή κόρη, λόγο των γενεθλίων της. Οπότε μια κούκλα Barbie είναι το τέλειο δώρο, όμως κατά την έξοδο από το εμπορικό ο μικρός αδερφός αρχίζει και παραπονιέται θέλει και εκείνος κάτι, ίσως ένα ζαχαρωτό να μπορούσε να τον ησυχάσει. Ίσως τα ζαχαρωτά να είναι για την μητέρα, δεδομένου ότι τα ψώνια την κούρασαν και χρειάζεται λίγη ενέργεια. Αυτά όλα τα σενάρια δείχνουν ότι τα ζαχαρωτά είναι μια παρορμητική αγορά που προστίθεται επάνω σε αυτήν της κούκλας Barbie.

Ωστόσο το αν το κατάστημα μπορεί να χρησιμοποιήσει αυτές τις πληροφορίες δεν είναι σαφές. Εντούτοις αυτός ο κανόνας προτείνει μια καλύτερη τοποθέτηση προϊόντων,

όπως για παράδειγμα να εξασφαλίσει ότι οι πελάτες πρέπει να περπατήσουν μέσω των διαδρόμων ζαχαρωτών στον γυρισμό τους από τα καταστήματα κούκλων. Μπορεί επίσης να προτείνει την σύνδεση και την προώθηση προϊόντων όπως την πώληση ζαχαρωτών και κούκλων από κοινού. Επίσης μπορεί να προτείνει ιδιαίτερους τρόπους να διαφημιστούν τα προϊόντα. Εντέλει επειδή ο κανόνας γίνεται κατανοητός εύκολα, προτείνει τις εύλογες αφορμές και τις πιθανές επεμβάσεις.

Τετριμμένοι κανόνες

Σύμφωνα με τους Berry και Linoff (2004) τα τετριμμένα αποτελέσματα είναι γνωστά ήδη από καθένα που είναι εξοικειωμένος με την επιχείρηση. Το δεύτερο παράδειγμα ("ότι οι πελάτες που αγοράζουν συμφωνίες συντήρησης είναι πολύ πιθανό να αγοράσουν μεγάλες συσκευές") είναι ένα παράδειγμα ενός τετριμμένου κανόνα. Στην πραγματικότητα, οι πελάτες αγοράζουν συγκεκριμένα συμφωνίες συντήρησης και τις μεγάλες συσκευές συγχρόνως. Για ποιόν λόγο άλλωστε θα αγόραζαν τις συμφωνίες συντήρησης; Συνήθως αυτά τα δύο διαφημίζονται μαζί, και πωλούνται σπάνια χωριστά (αν και όταν πωλείται κάτι χωριστά, αυτό είναι η μεγάλη συσκευή που πωλείται χωρίς τη συμφωνία παρά η συμφωνία που πωλείται χωρίς τη συσκευή). Ωστόσο αν και ο κανόνας ισχύει και υποστηρίζεται καλά από τα στοιχεία, παραμένει άχρηστος.

Επίσης ένα λεπτότερο πρόβλημα εμπίπτει στην ίδια κατηγορία. Το γεγονός ότι οι άνθρωποι που αγοράζουν την τριπλή κλήση ως τηλεφωνική υπηρεσία σχεδόν πάντα αγοράζουν και την αναμονή κλήσης μπορεί να είναι ένα αποτέλεσμα των προγραμμάτων μάρκετινγκ του παρελθόντος. Στην περίπτωση της επιλογής τηλεφωνικών υπηρεσιών, η τριπλή κλήση προσφέρεται μαζί με την αναμονή κλήσης, έτσι είναι δύσκολο να παραγγελθεί χωριστά. Κατ' επέκταση σε αυτήν την περίπτωση, η ανάλυση δεν παράγει αγώγιμα αποτελέσματα αλλά παράγει αποτελέσματα που είναι ήδη ενεργά. Αν και είναι ένας κίνδυνος για οποιαδήποτε τεχνική εξόρυξης δεδομένων, η ανάλυση καλαθιών αγοράς είναι ιδιαίτερα ευαίσθητη στην αναπαραγωγή επιτυχίας των προηγούμενων εκστρατειών μάρκετινγκ λόγω της εξάρτησής της από τα δεδομένα των σημείων πώλησης δηλαδή ακριβώς τα ίδια δεδομένα που καθορίζουν την επιτυχία της εκστρατείας. Εντέλει τα αποτελέσματα από την ανάλυση καλαθιών αγοράς μπορούν απλά να μετρούν την επιτυχία των προηγούμενων εκστρατειών μάρκετινγκ.

Ανεξήγητοι κανόνες

Τα ανεξήγητα αποτελέσματα είναι αποτελέσματα που φαίνονται να μην έχουν καμία εξήγηση και δεν προτείνουν κάποιο σχέδιο δράσης. Ο τρίτος κανόνας (Όταν ένα νέο κατάστημα ανοίγει, ένα από τα προϊόντα που αγοράζεται συχνότερα είναι τα καθαριστικά τουαλετών) προκαλεί ενδιαφέρον, βάζοντας σε πειρασμό με ένα νέο γεγονός αλλά παρέχοντας πληροφορίες που δεν δίνουν διορατικότητα στην καταναλωτική συμπεριφορά και τα εμπορεύματα ή δεν προτείνουν περαιτέρω ενέργειες.

3.3.3 Οικοδόμηση κανόνων ένωσης

Υπάρχουν τρία σημαντικά ζητήματα στη δημιουργία των κανόνων ένωσης:

- Επιλογή του σωστού συνόλου στοιχείων.
- Παραγωγή κανόνων με την αποκρυπτογράφηση.
- Υπερνίκηση των πρακτικών ορίων των υπεράριθμων στοιχείων.

Επιλογή του σωστού συνόλου στοιχείων

Τα στοιχεία που χρησιμοποιούνται για την εύρεση των κανόνων ένωσης είναι συγκεκριμένα τα δεδομένα της συναλλαγής στο σημείο της πώλησης. Επομένως η συγκέντρωση και η χρησιμοποίηση αυτών των στοιχείων είναι ένα κρίσιμο μέρος της εφαρμογής της ανάλυσης καλαθιών αγοράς, που εξαρτάται από τα στοιχεία που επιλέγονται για την ανάλυση.

Παραγωγή των κανόνων από όλα αυτά τα στοιχεία

Ο υπολογισμός του αριθμού εμφάνισης ενός συνδυασμού προϊόντων στα δεδομένα των συναλλαγών είναι ικανοποιητικός, αλλά ένας συνδυασμός προϊόντων δεν είναι ένας κανόνας. Συνήθως *ένας τέτοιος κανόνας έχει την ακόλουθη μορφή*: "Εάν ένας πελάτης αγοράσει το προϊόν Α, τότε ο πελάτης αναμένεται επίσης να αγοράσει το προϊόν Β."

Υπερνίκηση των πρακτικών ορίων

Η παραγωγή των κανόνων ένωσης είναι μια διαδικασία πολλαπλών βημάτων. Το πρόβλημα σε αυτήν την διαδικασία αν και δεν είναι προφανές δημιουργείτε όταν αυξάνεται ο αριθμός των προϊόντων στους συνδυασμούς κάτι που έχει ως αποτέλεσμα την απαίτηση περισσότερου υπολογισμού. Επομένως αυτό οδηγεί στην αύξηση του χρόνου υπολογισμού

και επίσης στην αύξηση του χρόνου αναμονής κατά την εξέταση των συνδυασμών που περιέχουν περισσότερα από τρία ή τέσσερα προϊόντα.

Η λύση σε αυτό το πρόβλημα είναι η περικοπή. Εντούτοις η περικοπή είναι μια τεχνική για την μείωση τον προϊόντων ανά συνδυασμό σε κάθε βήμα της εξέτασης των προϊόντων. Πιο συγκεκριμένα σε κάθε στάδιο της εξέτασης αυτός ο αλγόριθμος αποκλείει ορισμένους συνδυασμούς που δεν ικανοποιούν κάποιο κριτήριο.

3.3.4 Επέκταση των ιδεών

Οι βασικές ιδέες των κανόνων ένωσης μπορούν να έχουν διάφορες εφαρμογές, όπως την σύγκριση διαφόρων καταστημάτων και την βελτίωση του καθορισμού των κανόνων

Χρησιμοποίηση των κανόνων ένωσης για σύγκριση καταστημάτων

Η ανάλυση καλαθιών αγοράς χρησιμοποιείται συνήθως για να κάνει συγκρίσεις μέσα σε μια ενιαία αλυσίδα. Ωστόσο διαφορετικά καταστήματα έχουν και διαφορετικά σχέδια πωλήσεων, είτε για λόγους που έχουν να κάνουν με την περιοχή και τις συνήθειες, ή με την αποτελεσματικότητα της διαχείρισης του καταστήματος, ή λόγο του διαφορετικού τρόπου διαφήμισης παραδείγματος χάριν, τα κλιματιστικά μηχανήματα και οι ανεμιστήρες αγοράζονται συχνά κατά τη διάρκεια αύξησης της θερμοκρασίας, αλλά αυτή η αύξηση της θερμοκρασίας έχει επίδραση μόνο σε μια περιορισμένη περιοχή. Ένα άλλο παράδειγμα είναι, ότι σε μικρές περιοχές, τα δημογραφικά της περιοχής μπορούν να ασκήσουν πολύ μεγάλη επίδραση, παραδείγματος χάριν θα αναμέναμε τα καταστήματα που βρίσκονται σε πλούσιες περιοχές να ακολουθούν διαφορετικά σχέδια πωλήσεων από εκείνα σε φτωχότερες γειτονιές. Αυτά είναι παραδείγματα όπου η ανάλυση καλαθιών αγοράς μπορεί να βοηθήσει στην περιγραφή των διαφορών και να χρησιμεύσει ως ένα παράδειγμα χρησιμοποίησης της ανάλυσης καλαθιών αγοράς για την κατευθυνόμενη εξόρυξη δεδομένων.

Επειδή οι κανόνες ένωσης ανήκουν στην μη κατευθυνόμενη εξόρυξη δεδομένων, οι κανόνες ενεργούν ως αφετηρίες για περαιτέρω δοκιμαστικές υποθέσεις. Παραδείγματος χάριν γιατί ένα κατάστημα που δουλεύει εδώ και 5 χρόνια χρησιμοποιεί διαφορετικό σχέδιο πωλήσεων από ένα άλλο νέο κατάστημα. Χρησιμοποιώντας αυτήν την τεχνική, η ανάλυση καλαθιών αγοράς μπορεί να χρησιμοποιηθεί για πολλούς άλλους τύπους συγκρίσεων:

- Πωλήσεις κατά τη διάρκεια προωθήσεων σε σύγκριση με πωλήσεις σε άλλους χρόνους

- Σύγκριση πωλήσεων σε διάφορες γεωγραφικές περιοχές βάσει , νομού, της άμεσης περιοχής μάρκετινγκ (DMA), ή της χώρας
- Αστικές σε σύγκριση με προαστιακές πωλήσεις
- Εποχιακές διαφορές στα σχέδια πωλήσεων

Κανόνες διαχωρισμού

Ένας κανόνας διαχωρισμού είναι παρόμοιος με έναν κανόνα ένωσης με την μόνη διαφορά ότι μπορεί να χρησιμοποιεί το συνδετικό "και όχι" στον ορισμό εκτός από το συνδετικό "και." Ένας χαρακτηριστικός κανόνας διαχωρισμού μοιάζει με το παρακάτω παράδειγμα: *εάν Α και όχι Β, τότε Γ.*

Συγκεκριμένα οι κανόνες διαχωρισμού μπορούν να παραχθούν από μια απλή προσαρμογή του βασικού αλγορίθμου της ανάλυσης καλαθιών αγοράς. Η προσαρμογή γίνεται με την εισαγωγή ενός νέου συνόλου αντικειμένων που είναι τα αντίστροφα από τα αρχικά αντικείμενα. Έπειτα γίνεται, τροποποίηση κάθε συναλλαγής έτσι ώστε να περιέχει ένα αντίστροφο προϊόν εάν, και μόνο εάν, δεν περιέχει το αρχικό προϊόν.

Υπάρχουν τρία μειονεκτήματα στη συμπερίληψη αυτών των νέων στοιχείων. Κατ' αρχάς, διπλασιάζεται ο συνολικός αριθμός προϊόντων που χρησιμοποιούνται στην ανάλυση. Επίσης δεδομένου ότι το ποσό υπολογισμού αυξάνεται εκθετικά με τον αριθμό των προϊόντων, ο διπλασιασμός αυτός ρίχνει σοβαρά την απόδοση. Δεύτερον, το μέγεθος της συναλλαγής αυξάνεται επειδή περιλαμβάνει αυτά τα προϊόντα. Τέλος το τρίτο ζήτημα είναι ότι η συχνότητα των αντίστροφων προϊόντων τείνει να είναι πολύ μεγαλύτερη από τη συχνότητα των αρχικών προϊόντων. Έτσι, οι περιορισμοί τείνουν να παραγάγουν κανόνες στους οποίους όλα τα στοιχεία προβάλουν τον ακόλουθο κανόνα:
εάν ΟΧΙ Α και ΌΧΙ Β τότε ΌΧΙ Γ.
Εντέλει αυτοί οι κανόνες είναι λιγότερο πιθανό να είναι αγώγιμοι.

Διαδοχική ανάλυση χρησιμοποιώντας τους κανόνες ένωσης

Οι κανόνες ένωσης μπορούν να εντοπίζουν πράγματα που συμβαίνουν ταυτόχρονα όπως ποια προϊόντα αγοράζονται μία δεδομένη στιγμή. Επομένως θα πρέπει να εξεταστούν οι ακολουθίες γεγονότων και η σημασία τους. Μερικά παραδείγματα είναι:

- Οι ιδιοκτήτες ενός νέου σπιτιού αγοράζουν κουρτίνες πριν αγοράσουν τα έπιπλα.
- Οι πελάτες που αγοράζουν καινούργιες χορτοκοπτικές μηχανές είναι πολύ πιθανό να αγοράσουν μια νέα μάνικα ποτίσματος στις επόμενες 6 εβδομάδες.

Τα δεδομένα χρονικών σειρών απαιτούν συνήθως κάποιο τρόπο προσδιορισμού του πελάτη κατά τη διάρκεια του χρόνου. Παραδείγματος χάρη οι ανώνυμες συναλλαγές δεν μπορούν να αποκαλύψουν ότι οι νέοι ιδιοκτήτες σπιτιών αγοράζουν τις κουρτίνες τους προτού να αγοράσουν έπιπλα. Για να γίνει αυτό απαιτείται ο εντοπισμός κάθε πελάτη, καθώς επίσης και ποιοι πελάτες αγόρασαν πρόσφατα σπίτι. Δεδομένου ότι οι μεγάλες αγορές γίνονται συχνά με πιστωτικές κάρτες ή χρεωστικές κάρτες, αυτό δεν είναι πρόβλημα. Για τα προβλήματα σε άλλους τομής, όπως η έρευνα των επιπτώσεων ιατρικών θεραπειών ή της συμπεριφοράς πελατών σε μια τράπεζα, οι συναλλαγές τυπικά περιλαμβάνουν πληροφορίες ταυτότητας.(Parvatyar A. And Sheth J. N., 2001)

Η χρονική σειρά θα μπορούσαμε να πούμε πως είναι μια ακολουθία παραγγελμένων προϊόντων. Γενικά, η χρονική σειρά περιέχει προσδιοριστικές πληροφορίες για τον πελάτη, δεδομένου ότι αυτές οι πληροφορίες χρησιμοποιούνται για να συνδέσουν τις διαφορετικές συναλλαγές σε μια σειρά.

Προκειμένου να χρησιμοποιηθεί μια χρονική σειρά, τα στοιχεία συναλλαγής πρέπει να έχουν δύο πρόσθετα χαρακτηριστικά γνωρίσματα:

- Αλληλουχία των πληροφοριών έτσι ώστε να καθορίσει πότε οι συναλλαγές εμφανίστηκαν η μια σε σχέση με την άλλη
- Προσδιοριστικές πληροφορίες, όπως ο αριθμός λογαριασμού, η διεύθυνση, ή η ταυτότητα πελατών που προσδιορίζουν τις διαφορετικές συναλλαγές που ανήκουν στον ίδιο πελάτη ή την οικογένεια..

Η οικοδόμηση των διαδοχικών κανόνων είναι παρόμοια με αυτή των κανόνων ένωσης:

1. Όλα τα προϊόντα που αγοράζονται από έναν πελάτη αντιμετωπίζονται ως μια ενιαία παραγγελία, και κάθε προϊόν διατηρεί μία χρονική ‘‘σφραγίδα’’ που δείχνει πότε αγοράστηκε.
2. Η διαδικασία εύρεσης των ομάδων προϊόντων που εμφανίζονται μαζί είναι η ίδια
3. Για την ανάπτυξη των κανόνων, παίρνουμε υπόψη μόνο κανόνες όπου κάποια προϊόντα αγοράστηκαν πριν από κάποια άλλα. Εντέλει το αποτέλεσμα είναι ένα σύνολο κανόνων ένωσης που μπορεί να αποκαλύψει διαδοχικά σχέδια.

3.4 Κίνδυνοι και ανάλυση επιβίωσης (survival analysis and hazards)

Επιβίωση και κίνδυνοι είναι δύο όροι οι οποίοι χρησιμοποιούνται σε διάφορους τομείς. Στην πραγματικότητα αυτές οι δύο τεχνικές δεν συνδέονται συχνά με το μάρκετινγκ και αυτό γιατί οι όροι από μόνοι τους μας παραπέμπουν σε εικόνες τρόμου ή ακόμα και σε reality show. (Berry and Linoff, 2004)

Η ανάλυση επιβίωσης, που καλείται επίσης και ανάλυση του χρόνου γεγονότος, δεν είναι κάτι το οποίο θα έπρεπε να μας ανησυχήσει. Στην πραγματικότητα συμβαίνει το ακριβώς αντίθετο, η ανάλυση επιβίωσης είναι πολύ πολύτιμη για την κατανόηση των πελατών. Στην ουσία η επιβίωση μας ενημερώνει για το πότε οι πελάτες πρόκειται να κάνουν κάτι σημαντικό, όπως για παράδειγμα το τελείωμα της σχέσης τους με την επιχείρηση. Επίσης μας υποδεικνύει ποιοι παράγοντες συσχετίζονται με το γεγονός αυτό. Επίσης οι κίνδυνοι και οι καμπύλες επιβίωσης παρέχουν στιγμιότυπα των πελατών και των κύκλων ζωής τους, απαντώντας σε ερωτήσεις όπως: "Πόσο πρέπει να ανησυχούμε για το αν ένας πελάτης πρόκειται να αποχωρήσει στο εγγύς μέλλον;" ή "για το αν ένας πελάτης δεν έχει κάνει μια αγορά πρόσφατα και αν θα πρέπει να ανησυχούμε για το ότι ο πελάτης δεν θα επιστρέψει

Η προσέγγιση επιβίωσης είναι σκηνοθετημένη στη σημαντικότερη πλευρά της συμπεριφοράς πελατών, στην διάρκεια αξιώματος πελατών. Καταρχάς όλος αυτός ο χρόνος επαφής με τους πελάτες παρέχει έναν πλούτο πληροφοριών, ειδικά όταν σχετίζονται με ιδιαίτερα επιχειρησιακά προβλήματα. Το πόσο καιρό οι πελάτες θα παραμείνουν πελάτες είναι ένα μυστήριο, παρόλα αυτά είναι ένα μυστήριο που η συμπεριφορά πελατών στο παρελθόν μπορεί να φωτίσει. Πλέον σχεδόν κάθε επιχείρηση αναγνωρίζει την αξία της πίστης των πελατών και αυτό γιατί όσο περισσότερο οι πελάτες μένουν στην επιχείρηση, τόσο λιγότερο πιθανό είναι να αποχωρήσουν.

Στον περιβάλλον των πελατών, οι επιχειρήσεις έχουν να κάνουν με δεκάδες χιλιάδες στοιχεία, από την στιγμή που οι βάσεις δεδομένων πελατών περιέχουν συχνά στοιχεία που αφορούν εκατομμύρια πελατών όπως επίσης και των προηγούμενων πελατών. Επομένως ένα μεγάλο μέρος του στατιστικού υποβάθρου της ανάλυσης επιβίωσης στρέφεται στην εξαγωγή κάθε τελευταίας πληροφορίας από εκατοντάδες στοιχείων. Στις εφαρμογές εξόρυξης δεδομένων, οι όγκοι των στοιχείων είναι τόσο μεγάλοι που οι στατιστικές ανησυχίες για την ακρίβεια των αποτελεσμάτων αντικαθίστανται από ανησυχίες για τη διαχείριση των μεγάλων όγκων στοιχείων.

Η σημαντικότητα της ανάλυσης επιβίωσης είναι ότι παρέχει έναν τρόπο κατανόησης των χαρακτηριστικών του χρόνου- γεγονότος, όπως:

- Πότε ένας πελάτης είναι πιθανό να φύγει
- Πότε είναι πιθανό ένας πελάτης να μεταναστεύσει σε ένα νέο τμήμα πελατών
- Πότε είναι πιθανό ένας πελάτης να διευρύνει ή να 'στενέψει' την πελατειακή σχέση
- Τους παράγοντες στη σχέση πελατών που αυξάνουν ή μειώνουν την πιθανή διάρκεια αξιώματος
- Την ποσοτική επίδραση των διάφορων παραγόντων στη διάρκεια αξιώματος πελατών

Όλες αυτές οι ιδέες και τα χαρακτηριστικά που αφορούν τους πελάτες τροφοδοτούν άμεσα τη διαδικασία μάρκετινγκ με αποτέλεσμα να καθιστούν πιθανή την κατανόηση του χρονικού διαστήματος που οι διαφορετικές ομάδες πελατών είναι πιθανό να είναι κοντά στην επιχείρηση και ως εκ τούτου την πιθανότητα κερδοφορίας αυτών των τμημάτων. Επίσης καθιστούν πιθανό να προβλέψουν τους αριθμούς πελατών, που λαμβάνουν υπόψη μία νέα απόκτηση. Επίσης η ανάλυση επιβίωσης καθιστά πιθανό να καθορίσει ποιοι παράγοντες έχουν τη μεγαλύτερη επίδραση στην παραμονή των πελατών. Τέλος, η ανάλυση μπορεί να εφαρμοστεί σε πράγματα εκτός της διάρκειας αξιώματος πελατών, κάτι που καθιστά πιθανό να καθορίσει πότε κάποια γεγονότα, όπως εάν ένας πελάτης θα επιστρέψει σε μία ιστοσελίδα, είναι πιθανό να συμβεί. (Berry and Linoff, 2004)

3.4.1 Διατήρηση πελατών

Η διατήρηση πελατών είναι μια έννοια γνωστή στις περισσότερες επιχειρήσεις που ανησυχούν για τους πελάτες τους. Στην πραγματικότητα η διατήρηση είναι μια στενή προσέγγιση της επιβίωσης, ειδικά κατά την εξέταση μιας ομάδας πελατών που η συνεργασία όλων με την επιχείρηση ξεκίνα τον ίδιο σχεδόν χρόνο. Επίσης η διατήρηση παρέχει ένα γνωστό πλαίσιο για να εισαγάγει μερικές βασικές έννοιες της ανάλυσης επιβίωσης όπως την μέση διάρκεια ζωής πελατών και τον υπολογισμό του μέσου όρου περικομμένης διάρκειας αξιώματος πελατών.

Υπολογισμός της διατήρησης

Στην ουσία ο υπολογισμός της διατήρησης απαντά στην ερώτηση: Πόσο καιρό μένουν οι πελάτες στην επιχείρηση; Στην πραγματικότητα αυτή η φαινομενικά απλή ερώτηση γίνεται πιο περίπλοκη όταν εφαρμόζεται στον πραγματικό κόσμο. Ως εκ τούτου η κατανόηση της διατήρησης πελατών χρειάζεται δύο κομμάτια πληροφοριών:

- Πότε άρχισε ο κάθε πελάτης.
- Πότε σταμάτησε ο κάθε πελάτης.

Στην ουσία η διαφορά μεταξύ αυτών των δύο τιμών είναι η διάρκεια αξιώματος πελατών, ένα καλό μέτρο για την διατήρηση πελατών.

Σύμφωνα με τους Berrykαι Linoff (2004) οποιαδήποτε λογική βάση δεδομένων που ισχυρίζεται ότι αφορά τους πελάτες πρέπει να έχει αυτά τα στοιχεία προσιτά. Φυσικά, οι βάσεις δεδομένων μάρκετινγκ είναι σπάνια απλές. Εντούτοις υπάρχουν δύο προκλήσεις με αυτές τις έννοιες. Η πρώτη πρόκληση είναι η απόφαση σχετικά με τον ορισμό της έναρξης και της λήξης, μια απόφαση που εξαρτάται συχνά από τον τύπο της επιχείρησης και των διαθέσιμων στοιχείων. Η δεύτερη πρόκληση είναι τεχνική και αφορά την εύρεση αυτών των ημερομηνιών έναρξης και λήξης στα διαθέσιμα στοιχεία και την πιθανότητα να μην είναι προφανής.

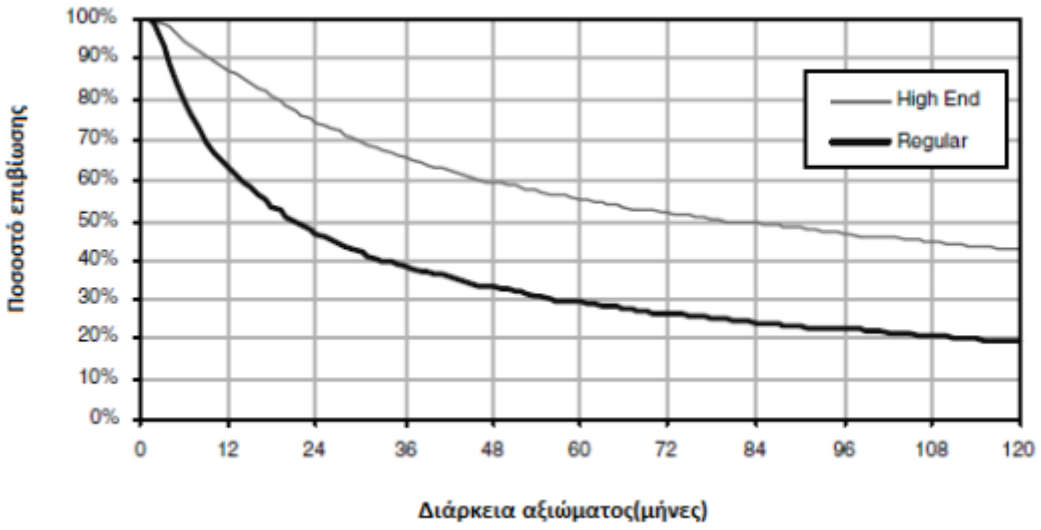
Τι αποκαλύπτει μια καμπύλη διατήρησης

Μόλις μπορούν να υπολογιστούν οι διάρκειες αξιώματος πελατών, μπορούν και να σχεδιαστούν σε μια καμπύλη διατήρησης, η οποία παρουσιάζει το ποσοστό των πελατών που διατηρούνται για μια συγκεκριμένη χρονική περίοδο. Στην πραγματικότητα αυτό είναι ένα συσσωρευτικό ιστόγραμμα, επειδή οι διάρκειες αξιώματος πελατών 3 μηνών συμπεριλαμβάνονται στις αναλογίες του 1 μήνα και των 2 μηνών. Ως εκ τούτου, μια καμπύλη διατήρησης αρχίζει πάντα από το 100 τοις εκατό.

Παραδείγματος χάριν αν υποθέσουμε ότι όλοι οι πελάτες αρχίζουν συγχρόνως, το σχήμα 3.2 συγκρίνει τη διατήρηση δύο ομάδων πελατών που άρχισαν σχεδόν στο ίδιο χρονικό σημείο, 10 έτη πριν. Επομένως τα σημεία στην καμπύλη παρουσιάζουν το ποσοστό των πελατών που διατηρήθηκαν για 1 έτος, για 2 έτη, και τα λοιπά. Εντούτοις μια τέτοια καμπύλη αρχίζει από το 100 τοις εκατό και κλίνει βαθμιαία προς τα κάτω. Όταν μια καμπύλη διατήρησης αντιπροσωπεύει πελάτες οι οποίοι αρχίζουν σχεδόν τον ίδιο χρόνο όπως σε αυτή την περίπτωση τότε έχουμε μια στενή προσέγγιση καμπύλης επιβίωσης.

Οι διαφορές στη διατήρηση μεταξύ των διαφορετικών ομάδων είναι σαφώς ορατές στο διάγραμμα. Αυτές οι διαφορές μπορούν να ποσολογηθούν. Ένα παράδειγμα ενός απλού μέτρου είναι να εξεταστεί η διατήρηση σε συγκεκριμένα χρονικά σημεία, παραδείγματος χάριν μετά από 10 έτη, το 24 τοις εκατό των αρχικών πελατών συνεχίζουν να είναι ακόμα πελάτες , και μόνο το ένα τρίτο αυτών συνεχίζουν να είναι μετά από5 έτη.

Ένας άλλος τρόπος σύγκρισης των διαφορετικών ομάδων είναι υπό την μορφή της εξής ερώτησης ‘πόσος είναι ο χρόνος που περνάει μέχρι την λήξη της συνεργασίας του μισού ποσοστού των πελατών ’δηλαδή ποια είναι η μεσαία διάρκεια ζωής πελατών. Συνοψίζοντας η διάμεσος είναι ένα χρήσιμο μέτρο επειδή ο μικρός αριθμός των πελατών που έχουν μακροχρόνια συνεργασία ή πολύ σύντομη δεν έχουν επιπτώσεις σε αυτό.



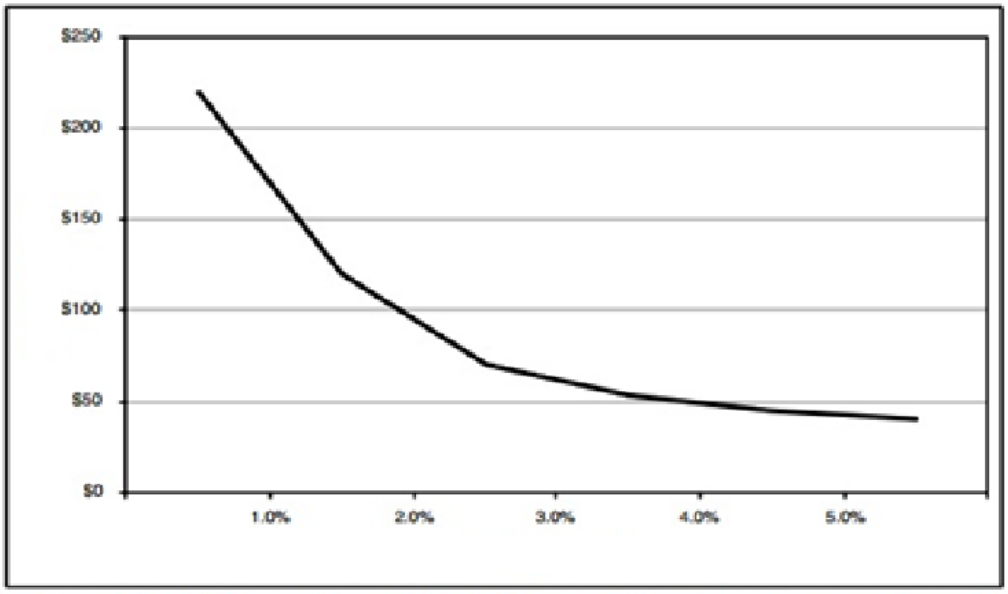
Σχέδιο 3.2 Η καμπύλη διατήρησης δείχνει πως οι ‘‘καλοί’’ πελάτες (High-end costumers) παραμένουν για μεγαλύτερο χρονικό διάστημα.

Γιατί παίζει ρόλο η αποχώρηση

Η αποχώρηση είναι σημαντική επειδή οι χαμένοι πελάτες πρέπει να αντικατασταθούν από νέους, και οι νέοι πελάτες είναι ακριβοί για να αποκτηθούν και παράγουν γενικά λιγότερα έσοδα βραχυπρόθεσμα από τους πελάτες που έχει ήδη η επιχείρηση. Αυτό ισχύει ιδιαίτερα στις μεγάλες βιομηχανίες όπου η αγορά είναι αρκετά κορεσμένη, κάποιος που πιθανώς να θελήσει το προϊόν ή η υπηρεσία πιθανώς ήδη την έχει από κάπου, έτσι η κύρια πηγή νέων πελατών είναι άνθρωποι που αφήνουν έναν ανταγωνιστή.

Το παρακάτω σχήμα διευκρινίζει ότι όσο η αγορά γίνεται πιο κορεσμένη και το ποσοστό ανταπόκρισης στις εκστρατείες απόκτησης μειώνεται, το κόστος για νέους πελάτες ανεβαίνει. Το διάγραμμα παρουσιάζει πόσο κοστίζει κάθε ένας νέος πελάτης για μια εκστρατεία άμεσου μάρκετινγκ μέσω ταχυδρόμησης, δεδομένου ότι οι δαπάνες αποστολής είναι $1 και περιλαμβάνει μια προσφορά $20 με κάποια μορφή, όπως ένα δελτίο ή ένα μειωμένο επιτόκιο σε μια πιστωτική κάρτα. Όταν το ποσοστό ανταπόκρισης στην εκστρατεία απόκτησης είναι υψηλό, όπως 5%, το κόστος ενός νέου πελάτη είναι $40. (Κοστίζει $100

δολάρια για να φθάσει σε 100 ανθρώπους, πέντε από τους οποίους ανταποκρίνονται με κόστος $20 δολαρίων κάθε ένα. Έτσι, πέντε νέοι πελάτες κοστίζουν $200 δολάρια.) Δεδομένου ότι το ποσοστό ανταπόκρισης μειώνεται, το κόστος αυξάνεται γρήγορα. Ώσπου να μειωθεί το ποσοστό ανταπόκρισης σε 1%, δαπάνες κάθε νέος πελάτης κοστίζει $200. Μέχρι κάποιο σημείο, έχει νόημα να ξοδευτούν εκείνα τα χρήματα για να διατηρήσει τους υπάρχοντες πελάτες παρά να προσελκύσει νέους.



Σχήμα 3.3 Η καμπύλη που δείχνει ότι όσο μειώνεται το ποσοστό ανταπόκρισης, τόσο αυξάνεται το κόστος απόκτησης νέων πελατών.

3.4.2 Κίνδυνοι

Όπως αναφερθήκαμε παραπάνω οι καμπύλες διατήρησης είναι πολύ χρήσιμες στην εξέταση και στην εξαγωγή συμπερασμάτων που έχουν να κάνουν με τους πελάτες. Γενικά αυτές οι καμπύλες είναι αρκετά απλές στην κατανόηση, αλλά μόνο από την άποψη των δεδομένων τους. Ωστόσο δεν υπάρχει καμία γενική μορφή, καμία παραμετρική μορφή, καμία μεγάλη θεωρία αποσύνθεσης πελατών. Δηλαδή το στοιχείο είναι και το μήνυμα.

Οι πιθανότητες κινδύνου επεκτείνουν αυτήν την ιδέα. Στην ουσία οι πιθανότητες κινδύνου, είναι ένα παράδειγμα μη παραμετρικής στατιστικής προσέγγισης αφήνοντας τα στοιχεία να δώσουν τα μηνύματα αντί της εύρεσης μιας ειδικής λειτουργίας η οποία θα επεξεργαστεί τα στοιχεία με αποτέλεσμα να προσφέρει κάποια μηνύματα. Συνοψίζοντας οι εμπειρικές πιθανότητες κινδύνου αφήνουν απλά τα ιστορικά στοιχεία να καθορίσουν αυτό

που είναι πιθανό να συμβεί, χωρίς προσπάθεια εγκατάστασης των στοιχείων σε κάποια προδικασμένη μορφή. Παρέχουν επίσης τη διορατικότητα στη διατήρηση πελατών και καθιστούν πιθανό να παραγάγουν μία 'καθαρή' καμπύλη διατήρησης η οποία ονομάζεται καμπύλη επιβίωσης.

Η βασική ιδέα της πιθανότητας κινδύνου

Σύμφωνα με τους Berry και Linoff (2004) μια πιθανότητα κινδύνου απαντά στην ακόλουθη ερώτηση:

Αν υποθέσουμε ότι ένας πελάτης έχει επιζήσει για ένα ορισμένο χρονικό διάστημα, έτσι η διάρκεια αξιώματος του πελάτη είναιt. Ποια είναι η πιθανότητα ο πελάτης να αποχωρίσει πριν από το t+1; Ένας άλλος τρόπος να το εκφράσουμε είναι: ότι ο κίνδυνος στο χρόνο t είναι ο κίνδυνος να χαθούν πελάτες μεταξύ του χρόνου t και του χρόνου t+1. Για να γίνουν κατανοητές οι πιθανότητες κινδύνου θα χρησιμοποιηθεί ένα παράδειγμα των κινδύνων που δεν αφορά τον κόσμο των επιχειρήσεων, ας εξετάσουμε τους πίνακες ζωής, οι οποίοι περιγράφουν την πιθανότητα κάποιου να πεθάνει σε μια συγκεκριμένη ηλικία. Ο πίνακας 3.3 παρουσιάζει αυτά τα στοιχεία, για τον αμερικανικό πληθυσμό το 2000:

ΗΛΙΚΙΑ ΠΛΥΘΙΣΜΟΥ	ΠΟΣΟΣΤΟ ΘΝΗΣΙΜΟΤΗΤΑΣ ΠΛΥΘΙΣΜΟΥ
0-1 Χρόνια	0,73%
1-4 Χρόνια	0,03%
5-9 Χρόνια	0,02%
10-14 Χρόνια	0,02%
15-19 Χρόνια	0,07%
20-24 Χρόνια	0,10%
25-29 Χρόνια	0,10%
30-34 Χρόνια	0,12%
35-39 Χρόνια	0,16%
40-44 Χρόνια	0,24%
45-49 Χρόνια	0,36%
50-54 Χρόνια	0,52%
55-59 Χρόνια	0,80%
60-64 Χρόνια	1,26%
65-69 Χρόνια	1,93%
70-74 Χρόνια	2,97%
75-79 Χρόνια	4,56%
80-84 Χρόνια	7,40%
85+ Χρόνια	15.32%

Πίνακας 3.1 Ο κίνδυνος θανάτου κατά την περίοδο 2000-2001 στης Ηνωμένες Πολιτείες παρουσιασμένος ως πίνακας ζωης

Ένας πίνακας ζωής είναι ένα καλό παράδειγμα των κινδύνων. Ως εκ τούτου η πιθανότητα ενός νήπιου να πεθάνει πριν από τα πρώτα γενέθλιά του είναι 1 στις 137. Το ποσοστό θνησιμότητας πέφτει κατακόρυφα έπειτα, αλλά τελικά αυξάνεται σταθερά. Μόνο όταν κάποιος είναι περίπου 55 ετών αυξάνεται ο κίνδυνος σε ποσοστό ανάλογο αυτού του πρώτου έτους.

Αυτό είναι μια χαρακτηριστική μορφή κινδύνου που καλείται μορφή ‘μπανιέρας’. Πιο συγκεκριμένα οι κίνδυνοι αρχίζουν με υψηλά ποσοστά, πέφτουν και παραμένουν χαμηλά για πολύ, και έπειτα βαθμιαία αυξάνονται και πάλι.

Η ίδια ιδέα μπορεί να εφαρμοστεί και στη διάρκεια αξιώματος πελατών, αν και οι κίνδυνοι πελατών υπολογίζονται συγκεκριμένα με βάση την ημέρα, την εβδομάδα, ή το μήνα αντί το έτος. επίσης ο υπολογισμός ενός κινδύνου για μια δεδομένη διάρκεια αξιώματος t απαιτεί μόνο δύο κομμάτια στοιχείων. Το πρώτο είναι ο αριθμός πελατών που σταμάτησαν στον χρόνο t (ή μεταξύ του t και t+1).Το δεύτερο είναι ο συνολικός αριθμός πελατών που θα μπορούσαν να έχουν σταματήσει κατά τη διάρκεια αυτής της περιόδου. Εντούτοις αυτό αποτελείται από όλους τους πελάτες των οποίων η διάρκεια αξιώματος είναι μεγαλύτερη ή ίση με t, συμπεριλαμβανομένων και εκείνων που σταμάτησαν στο χρόνο t. Σύμφωνα με αυτά η πιθανότητα κινδύνου είναι η αναλογία αυτών των δύο αριθμών, και όντας μια πιθανότητα, ο κίνδυνος είναι πάντα μεταξύ 0 και 1. Τέλος αυτοί οι υπολογισμοί κινδύνου παρέχονται από λειτουργίες ζωής σε στατιστικά λογισμικά όπως είναι το SAS και το SPSS.

Παραδείγματα των λειτουργιών κινδύνου

Στην συνέχεια αναφέρουμε μερικά παραδείγματα των λειτουργιών κινδύνου. Αυτά τα παραδείγματα μπορούν να βοηθήσουν στην κατανόηση τους, μέσο της εξέτασης των πιθανοτήτων κινδύνου. Τα πρώτα δύο παραδείγματα είναι βασικά και το τρίτο εξηγεί το πώς οι κίνδυνοι μπορούν να χρησιμοποιηθούν για να παρέχουν μια λεπτομερή εικόνα της διάρκειας ζωής των πελατών.

Σταθερός κίνδυνος

Στην ουσία αυτό που λέει ο σταθερός κίνδυνος είναι ότι ο κίνδυνος της αναχώρησης πελατών είναι ακριβώς ίδιος, ανεξάρτητα από το πόσο καιρό οι πελάτες είναι στην επιχείρηση. Στην γραφική παράσταση ο σταθερός κίνδυνος μοιάζει με μια οριζόντια γραμμή.

Εάν θεωρήσουμε πως ο κίνδυνος μετριέται σε μέρες, και είναι ένα σταθερό 0,1%, δηλαδή ότι ένας πελάτης στους χίλιους αποχωρεί κάθε ημέρα, μετά από ένα έτος (365 ημέρες), σημαίνει ότι περίπου 30,6% των πελατών θα έχουν φύγει. Εν συνεχεία θα χρειαστούν περίπου 692 ημέρες για να αποχωρίσουν οι μισοί από αυτούς και άλλες 692 ημέρες για τους μισούς από αυτούς, και συνεχίζετε αυτή η διαδικασία

Ουσιαστικά ο σταθερός κίνδυνος σημαίνει ότι η πιθανότητα μιας αποχώρησης πελατών δεν ποικίλλει ανάλογα με το χρονικό διάστημα που ο πελάτης βρίσκεται στην επιχείρηση. Στην πραγματικότητα, ένας σταθερός κίνδυνος διατήρησης θα προσαρμοζόταν σε μια εκθετική μορφή για την καμπύλη διατήρησης.

Κίνδυνος με μορφή 'μπανιέρα'

Όπως αναφέρθηκε παραπάνω ο πίνακας ζωής για τον αμερικάνικο πληθυσμό ήταν ένα παράδειγμα για την λειτουργία κινδύνου και αυτό που τον χαρακτηρίζει είναι ότι ο κίνδυνος με μορφή μπανιέρας αρχίζει με ένα υψηλό ποσοστό, κατόπιν μειώνετε ώσπου τα ποσοστά μένουν σταθερά για ένα μεγάλο χρονικό διάστημα , και τελικά σημειώνεται αύξηση του κινδύνου και πάλι.

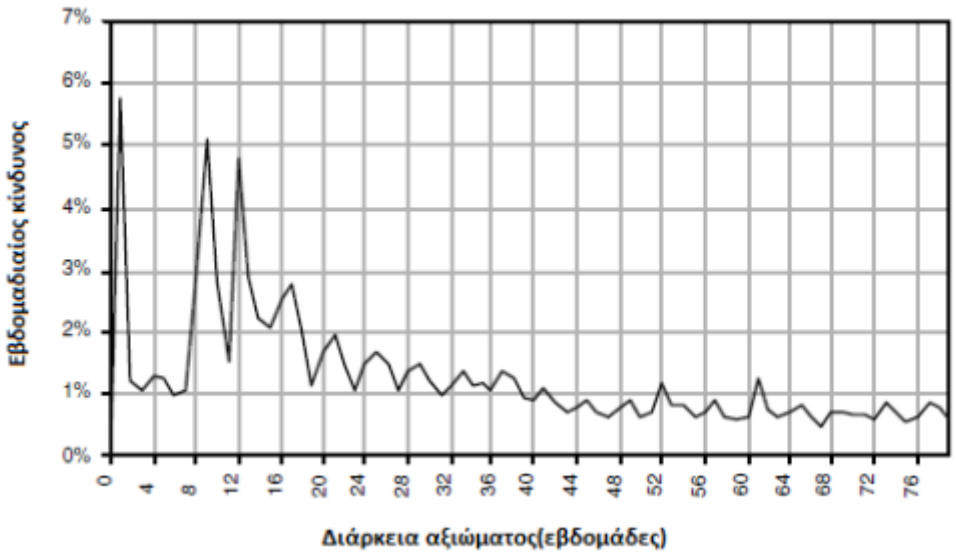
Ένα παράδειγμα του επιχειρηματικού κόσμου είναι το φαινόμενο που προκαλείτε από τους πελάτες που υπογράφουν συμβάσεις για διάστημα μεγαλύτερο του ενός χρόνου (παραδείγματος χάριν, για τα κινητά τηλέφωνα ή τις ISP υπηρεσίες). Συνήθως νωρίς μετά την σύμβαση οι πελάτες αποχωρούν επειδή η υπηρεσία δεν είναι αντιπροσωπευτική ή επειδή απλά δεν επιθυμούν να πληρώνουν. Επίσης κατά τη διάρκεια της περιόδου της σύμβασης, οι πελάτες αποφεύγουν την ακύρωση, είτε λόγω των οικονομικών ποινικών ρητρών είτε ίσως λόγω ενός συναισθήματος υποχρέωσης προς τους όρους της αρχικής σύμβασης.

Μόλις λήξει η σύμβαση οι πελάτες συνήθως βιάζονται να αποχωρήσουν και το ποσοστό αυτό αυξάνετε από την στιγμή που οι πελάτες έχουν ελευθερωθεί από την σύμβαση, μπορεί βέβαια να υπάρξουν και άλλοι λόγοι, όπως το ότι το προϊόν ή η υπηρεσία δεν έχει διατιμηθεί ανταγωνιστικά, κάτι το οποίο αναγκάζει τους πελάτες να αποχωρήσουν. Ως εκ τούτου η αγορά αλλάζει και οι πελάτες ανταποκρίνονται σε αυτές τις αλλαγές. Έτσι δεδομένου ότι οι τηλεφωνικές δαπάνες μειώνονται, οι πελάτες είναι πιθανότερο να μετακινηθούν σε έναν ανταγωνιστή από το να διαπραγματευτούν με τον τρέχοντα προμηθευτή τους για μία χαμηλότερη τιμή.

Το παράδειγμα κινδύνου μιας συνδρομητικής επιχείρησης

Στην συνέχεια θα χρησιμοποιήσουμε το παράδειγμα μιας λειτουργίας κινδύνου, για μια επιχείρηση που πωλεί μια υπηρεσία βασισμένη στην συνδρομή, για να δείξουμε το πώς οι κίνδυνοι μπορούν να χρησιμοποιηθούν για να παρέχουν μια λεπτομερή εικόνα της διάρκειας ζωής των πελατών. Πιο συγκεκριμένα αυτή η λειτουργία κινδύνου μετρά την πιθανότητα ενός πελάτη να αποχωρήσει μετά από έναν δεδομένο αριθμό εβδομάδων από την εγγραφή.

Εξετάζοντας την καμπύλη εντοπίζουμε διάφορα ενδιαφέροντα χαρακτηριστικά. Καταρχάς όπως φαίνεται αρχίζει με ένα υψηλό ποσοστό κινδύνου. Αυτοί είναι πελάτες οι οποίοι εγγράφονται, αλλά δεν μπορούν να γίνουν δέκτες της υπηρεσίας για κάποιο τεχνικό λόγο όπως την μη έγκριση της πιστωτικής κάρτας .Επίσης σε μερικές περιπτώσεις, οι πελάτες δεν συνειδητοποιούν ότι έχουν εγγραφεί.



Σχέδιο 3.4 Σχεδιάγραμμα πιθανοτήτων κινδύνου μιας συνδρομητικής επιχείρησης.

Έπειτα, υπάρχει χαρακτηριστικό γνώρισμα σε σχήμα Μ, με τις αιχμές στην 9η και 11η εβδομάδα αντίστοιχα. Η πρώτη αιχμή, εμφανίζεται στους 2 μήνες λόγω της μη πληρωμής. Συγκεκριμένα αυτό συμβαίνει διότι οι πελάτες που δεν πληρώνουν τους λογαριασμούς, ή που ακυρώνουν τις δαπάνες πιστωτικών καρτών, ακυρώνονται λόγο μη πληρωμής μετά από περίπου 2 μήνες. Δεδομένου ότι ένας σημαντικός αριθμός πελατών φεύγει εκείνη την στιγμή, το ποσοστό κινδύνου αυξάνεται.

Η δεύτερη αιχμή στο "Μ" συμπίπτει με το τέλος της αρχικής προώθησης που προσφέρει μία ειδική τιμολόγηση. Αυτή η προώθηση διαρκεί περίπου 3 μήνες, και έπειτα οι

πελάτες πρέπει να πληρώνουν την πλήρη τιμή. Επομένως πολλοί αποφασίζουν ότι δεν έχουν ανάγκη την υπηρεσία. Επίσης είναι δυνατό πολλοί από αυτούς τους πελάτες να επανεμφανιστούν για να εκμεταλλευθούν άλλες προωθήσεις.

Μετά από τους πρώτους 3 μήνες, η λειτουργία κινδύνου δεν εμφανίζει άλλα υψηλά ποσοστά κινδύνου. Βέβαια υπάρχει ένα μικρό σκαμπανέβασμα των ποσοστών, κάθε 4 ή 5 εβδομάδες. το οποίο αντιστοιχεί στο μηνιαίο κύκλο τιμολόγησης. Αυτό δείχνει ότι οι πελάτες είναι πιθανότερο να σταματήσουν αμέσως αφότου λαμβάνουν έναν λογαριασμό.

Το διάγραμμα επίσης δείχνει ότι υπάρχει μια πτώση στο ποσοστό κινδύνου μακροχρόνια. Αυτή η πτώση είναι θετική, δεδομένου ότι σημαίνει ότι όσο περισσότερο οι πελάτες παραμένουν στην εταιρεία, είναι λιγότερο πιθανό να αποχωρήσουν. Δηλαδή ότι οι πελάτες γίνονται περισσότερο πιστοί όσο παραμένουν στην επιχείρηση.

3.4.3 Από τους κινδύνους στην επιβίωση

Από τις λειτουργίες κινδύνου, είναι δυνατό να δημιουργηθεί μια πολύ παρόμοια καμπύλη, αποκαλούμενη καμπύλη επιβίωσης. Η καμπύλη επιβίωσης είναι πιο χρήσιμη και υπό πολλές έννοιες ακριβέστερη.

Επιβίωση

Όπως αναφέρθηκε και παραπάνω οι κίνδυνοι δίνουν την πιθανότητα ενός πελάτη να αποχωρήσει σε ένα συγκεκριμένο χρονικό σημείο. Η επιβίωση, αφ' ετέρου, δίνει την πιθανότητα ενός πελάτη να συντηρηθεί μέχρι εκείνο τον χρόνο. Ωστόσο οι τιμές επιβίωσης υπολογίζονται άμεσα από τους κινδύνους. Πιο συγκεκριμένα σε οποιοδήποτε χρονικό σημείο, η πιθανότητα ενός πελάτη να επιζεί στην επόμενη μονάδα του χρόνου είναι (1 – κίνδυνος), οι οποία ονομάζεται επιβίωση στο χρόνο t. Εντούτοις για τον υπολογισμό της πλήρης επιβίωσης σε μία δεδομένη στιγμή απαιτούνται όλες οι τιμές επιβίωσης μέχρι εκείνο το σημείο έτσι ώστε να πολλαπλασιαστούν μεταξύ τους. Επίσης η αξία επιβίωσης είναι 1 (ή 100 %) στο χρόνο 0, δεδομένου ότι όλοι οι πελάτες που περιλαμβάνονται στην ανάλυση επιζούν στην αρχή της ανάλυσης.

Από τη στιγμή που ο κίνδυνος είναι πάντα μεταξύ 0 και 1, η επιβίωση είναι επίσης μεταξύ 0 και 1. Ως εκ τούτου, η τιμή της επιβίωσης μειώνεται επειδή κάθε διαδοχική αξία πολλαπλασιάζεται με έναν αριθμό μικρότερο από το 1. Επομένως η καμπύλη επιβίωσης ξεκινάει από την τιμή 1, και παίρνει μια ήπια κατηφορική πορεία, ίσως μερικές φορές να γίνεται επίπεδη, αλλά ποτέ δεν παίρνει κλήση προς τα επάνω.

Οι καμπύλες επιβίωσης έχουν περισσότερο νόημα στην διατήρηση πελατών από τις καμπύλες διατήρησης αυτό συμβαίνει διότι η καμπύλη επιβίωσης είναι ομαλότερη, και γιατί έχει πάντα κλίση προς τα κάτω σε αντίθεση με την καμπύλη διατήρησης οι οποία παρουσιάζει σκαμπανεβάσματα στο διάγραμμα.
Η διαφορά μεταξύ της καμπύλης διατήρησης και της καμπύλης επιβίωσης είναι ότι η καμπύλη διατήρησης συγχωνεύει πολλές εικόνες διαφορετικών πελατών από το παρελθόν.

Εντούτοις, στις καμπύλες διατήρησης, οι που πελάτες ξεκινούν την συνεργασία τους με την επιχείρηση σε διαφορετικά χρονικά σημεία έχουν και διαφορετικές προοπτικές. Επίσης οποιοδήποτε δεδομένο σημείο στην καμπύλη διατήρησης είναι κοντά στην πραγματική αξία διατήρησης παρόλα αυτά, εξετάζοντας την ως σύνολο, φαίνεται ανομοιόμορφη. Ένας τρόπος να αφαιρεθεί αυτή οι ανομοιομορφία είναι να γίνει εστίαση στους πελάτες που αρχίζουν σχεδόν τον ίδιο χρόνο, όπως αναφέρθηκε νωρίτερα. Εντούτοις, αυτό μειώνει πολύ το ποσό στοιχείων που συμβάλλουν στην καμπύλη.

Η καμπύλη επιβίωσης, αφ' ετέρου, εξετάζει όσο το δυνατόν περισσότερους πελάτες, όχι μόνο αυτούς που ξεκίνησαν μια συγκεκριμένη χρονική περίοδο. Επομένως η επιβίωση σε οποιοδήποτε δεδομένο χρονικό σημείο t χρησιμοποιεί πληροφορίες από όλους τους πελάτες. Εντούτοις η επιβίωση, υπολογίζεται με το συνδυασμό όλων των πληροφοριών που αφορούν τους κινδύνους.

Κατά την ανάλυση των πελατών, οι κίνδυνοι και η επιβίωση παρέχουν πολύτιμες πληροφορίες σχετικά με τους πελάτες. Επειδή η επιβίωση είναι συσσωρευτική, δίνει μια καλή συνοπτική αξία για τη σύγκριση των διαφορετικών ομάδων πελατών. Επίσης η επιβίωση χρησιμοποιείται για τον υπολογισμό της *μεσαίας διάρκειας ζωής* πελατών και τη μικρότερη διάρκεια αξιώματος πελατών, που τροφοδοτεί στη συνέχεια άλλους υπολογισμούς, όπως την αξία πελατών.

Επειδή όπως αναφέρθηκε η επιβίωση είναι συσσωρευτική, είναι δύσκολο να φανούν τα σχέδια σε ένα ιδιαίτερο χρονικό σημείο. Σε αντίθεση οι κίνδυνοι καθιστούν αυτές τις συγκεκριμένες αιτίες προφανέστερες. Πολλές φορές είναι δυνατό να προσδιοριστούν γεγονότα κατά τη διάρκεια του κύκλου ζωής πελατών τα οποία οδηγούν σε κινδύνους. Εν κατακλείδι οι καμπύλες επιβίωσης δεν δίνουν έμφαση σε τέτοια γεγονότα όπως κάνουν οι κίνδυνοι.

Ανάλυση επιβίωσης στην πράξη

Η ανάλυση επιβίωσης έχει αποδειχθεί πολύ πολύτιμη για την κατανόηση των πελατών και την αξιολόγηση των προσπαθειών μάρκετινγκ από την άποψη της διατήρησης πελατών.

Επίσης παρέχει τον τρόπο με τον οποίο μπορούμε να υπολογίσουμε τον χρόνο μέχρι την εμφάνιση ενός προβλήματος.

Χειρισμός των διαφορετικών τύπων τριβών

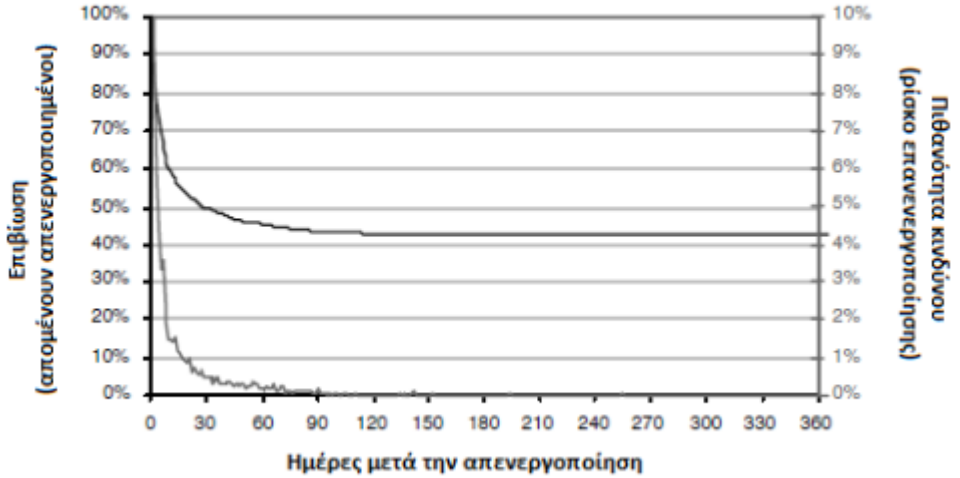
Οι επιχειρήσεις πρέπει να εξετάζουν τους πελάτες που φεύγουν για ποικίλους λόγους. Νωρίτερα, περιγράψαμε τις πιθανότητες κινδύνου και εξηγήσαμε πώς οι κίνδυνοι επεξηγούν τις πτυχές της επιχείρησης που έχουν επιπτώσεις στον κύκλο ζωής πελατών. Ειδικότερα, όπως παρατηρήθηκε στο διάγραμμα κινδύνου οι αιχμές συνέπεσαν με τις επιχειρησιακές διαδικασίες που ανάγκασαν τους πελάτες να αποχωρήσουν επειδή δεν πλήρωναν τους λογαριασμούς τους. Δεδομένου ότι αυτοί οι πελάτες αντιμετωπίζονται διαφορετικά, θα μπορούσαν να αφαιρεθούν εξ ολοκλήρου από τον υπολογισμό κινδύνου. Παρόλα αυτά αυτό είναι μια λανθασμένη προσέγγιση. Το πρόβλημα σε αυτό είναι ότι η γνώση των πελατών που πρέπει να αφαιρεθούν αποκτάται εφόσον οι πελάτες έχουν αναγκαστεί να αποχωρήσουν. Δεν είναι καλό να χρησιμοποιηθεί τέτοια γνώση, που αποκτιέται στο τέλος της σχέσης πελατών, για το φιλτράρισμα των πελατών προς ανάλυση.

Η σωστή προσέγγιση είναι η διάσπαση αυτού σε δύο προβλήματα. Δηλαδή ποιοι είναι οι κίνδυνοι της εθελοντικής τριβής; Και ποιοι είναι οι κίνδυνοι της εξαναγκασμένης τριβής; Τα δύο αυτά προβλήματα χρησιμοποιούν όλους τους πελάτες, ελέγχοντας τους πελάτες που φεύγουν λόγω άλλων παραγόντων. Κατά τον υπολογισμό των κινδύνων για την εθελοντική τριβή, όταν ένας πελάτης αναγκάζεται να φύγει, τότε ο πελάτης συμπεριλαμβάνεται στην ανάλυση και ελέγχετε έως ότου το σημείο που αποχωρεί. Αυτό συμβαίνει διότι μέχρι το σημείο όπου αναγκάστηκε να φύγει ο πελάτης δεν έφυγε εθελοντικά. (Daniel T. Larose, 2006)

Πότε ένας πελάτης θα επιστρέψει

Η ανάλυση επιβίωσης μπορεί να χρησιμοποιηθεί για πολλά πράγματα εκτός από την πρόβλεψη της πιθανότητας να συμβεί κάτι κακό. Παραδείγματος χάριν, η ανάλυση επιβίωσης μπορεί να χρησιμοποιηθεί για να υπολογίσει πότε οι πελάτες θα επιστρέψουν αφότου έχουν αποχωρήσει.

Το σχήμα 3.5 παρουσιάζει μια καμπύλη επιβίωσης και τους κινδύνους για την επανενεργοποίηση των πελατών αφότου απενεργοποιήσουν την υπηρεσία κινητών τηλεφώνων τους. Σε αυτήν την περίπτωση, ο κίνδυνος είναι η πιθανότητα ότι ένας πελάτης επιστρέφει έναν δεδομένο αριθμό ημερών μετά από την απενεργοποίηση.



Σχήμα 3.5 Καμπύλη επιβίωσης (πάνω) και κινδύνου (κάτω) για την επανενεργοποίηση πελατών κινητής τηλεφωνίας.

Υπάρχουν διάφορα ενδιαφέροντα χαρακτηριστικά γνωρίσματα σε αυτές τις καμπύλες. Κατ' αρχάς, το αρχικό ποσοστό επανενεργοποίησης είναι πολύ υψηλό. Πιο συγκεκριμένα την πρώτη εβδομάδα, περισσότερο από το ένα τρίτο των πελατών επανενεργοποιεί. Οι επιχειρησιακοί κανόνες μπορούν αν εξηγήσουν αυτό το φαινόμενο. Πολλές απενεργοποιήσεις οφείλονται σε πελάτες που δεν πληρώνουν τους λογαριασμούς τους. Πολλοί από αυτούς τους πελάτες περιμένουν μέχρι την τελευταία στιγμή για να πληρώσουν, στην πραγματικότητα σκοπεύουν να κρατήσουν τα τηλέφωνά τους απλώς δεν τους αρέσει να πληρώνουν το λογαριασμό. Εντούτοις, μόλις σταματήσει η λειτουργία του τηλέφωνου, πληρώνουν αμέσως. Μετά από 90 ημέρες, οι κίνδυνοι είναι ουσιαστικά μηδέν δηλαδή οι πελάτες δεν επανενεργοποιούν. Εντούτοις άλλη μια φορά, οι επιχειρησιακές διαδικασίες παρέχουν οδηγίες. Οι αριθμοί τηλεφώνου διατηρούνται για 90 ημέρες αφότου αποχωρήσουν οι πελάτες. Συνήθως, όταν επανενεργοποιούν οι πελάτες, θέλουν να κρατήσουν τον ίδιο αριθμό τηλεφώνου. Ωστόσο μετά από 90 ημέρες, ο αριθμός μπορεί να έχει επανεκχωρηθεί, και ο πελάτης θα πρέπει να πάρει έναν νέο αριθμό τηλεφώνου.

Στην παραπάνω περίπτωση, η ανάλυση χρησιμοποίησε τους αριθμούς τηλεφώνου σε συνδυασμό με μια ταυτότητα λογαριασμού. Αυτό λίγο πολύ εγγυήθηκε ότι η αντιστοιχία ήταν ακριβής, δεδομένου ότι οι επανενεργοποιημένοι πελάτες διατήρησαν τους αριθμούς τηλεφώνου και τις πληροφορίες τιμολόγησής τους. Αυτό είναι πολύ συντηρητικό αλλά λειτουργεί για την εύρεση των επανενεργοποιήσεων. Ωστόσο δεν λειτουργεί για την εύρεση

άλλων τύπων επιστροφής πελατών, όπως για παράδειγμα πελάτες που είναι πρόθυμοι να ανακυκλώνουν αριθμούς τηλεφώνου προκειμένου να αποκτήσουν κάποιες εκπτώσεις.

Μια άλλη προσέγγιση είναι η προσπάθεια προσδιορισμού ατόμων κατά τη διάρκεια του χρόνου, ακόμα και όταν αυτοί έχουν διαφορετικούς λογαριασμούς. Για τις επιχειρήσεις που συλλέγουν αριθμούς κοινωνικής ασφάλισης ή τους αριθμούς άδειας οδήγησης, τέτοιοι προσδιοριστικοί αριθμοί μπορούν να συνδέσουν λογαριασμούς κατά τη διάρκεια του χρόνου. Μερικές φορές είναι αρκετό για σκοπούς σύνδεσης το ταίριασμα ονομάτων, διευθύνσεων, αριθμών τηλεφώνου, ή και πιστωτικών καρτών. Αυτή η σύνδεση παρέχει πληροφορίες που απαιτούνται για να προσδιοριστούν ποιοι είναι νέοι πελάτες και ποιοι είναι στην πραγματικότητα προηγούμενοι πελάτες που έχουν κερδηθεί πίσω. (Berry and Linoff, 1999)

Μια καλή πτυχή της χρησιμοποίησης της ανάλυσης επιβίωσης είναι ότι είναι εύκολο να κάνουμε ερωτήσεις που αφορούν αποτελέσματα διαφορετικών αρχικών καταστάσεων όπως τον αριθμό των επισκέψεων ενός πελάτη στην επιχείρηση κατά το παρελθόν. Επίσης χρησιμοποιώντας τους ανάλογους κινδύνους, είναι δυνατό να καθοριστούν ποιες μεταβλητές έχουν την περισσότερη επίδραση στο επιθυμητό αποτέλεσμα, που περιλαμβάνει ποιες επεμβάσεις είναι περισσότερο και ποιες λιγότερο πιθανό να έχουν επιτυχία.

Πρόβλεψη

Μια άλλη ενδιαφέρουσα εφαρμογή της ανάλυσης επιβίωσης είναι η πρόβλεψη του μελλοντικού αριθμού πελατών, ή του αριθμού των αποχωρήσεων από την επιχείρηση για μια δεδομένη στιγμή στο μέλλον. Συγκεκριμένα, η επιβίωση εκτελεί μια εργασία υπολογισμού του αριθμού πελατών που θα προσκολληθούν στην επιχείρηση ένα δεδομένο χρονικό διάστημα.

Υπάρχουν δύο συστατικά σε οποιαδήποτε τέτοια πρόβλεψη. Το πρώτο είναι ένα πρότυπο των υπαρχόντων πελατών, το οποίο μπορεί να λάβει υπόψη τις διάφορες μεταβλητές κατά τη διάρκεια του κύκλου ζωής του πελάτη. Ένα τέτοιο πρότυπο λειτουργεί με την εφαρμογή ενός ή περισσότερων μοντέλων επιβίωσης σε όλους τους πελάτες. Εντούτοις εάν ένας πελάτης έχει επιζήσει για 100 ημέρες, τότε η πιθανότητα να αποχωρήσει την επόμενη μέρα είναι ο κίνδυνος την ημέρα 100.

Το δεύτερο συστατικό της πρόβλεψης επιπέδου πελατών είναι λίγο δυσκολότερο να υπολογιστεί. Αυτό το συστατικό είναι η επίδραση των νέων πελατών στην πρόβλεψη, και η δυσκολία δεν είναι τεχνική. Η πρόκληση είναι οι εκτιμήσεις για τις νέες ενάρξεις πελατών. Ευτυχώς, υπάρχουν συχνά προβλέψεις προϋπολογισμών που περιέχουν τις νέες ενάρξεις, κατανεμημένες ανά προϊόν, κανάλι, ή και γεωγραφικά. Φυσικά, η πρόβλεψη είναι τόσο

ακριβής όσο ο προϋπολογισμός. Εν τούτοις, η πρόβλεψη, βασισμένη σε τεχνικές επιβίωσης, μπορεί να ενσωματωθεί στη διαδικασία διαχείρισης πραγματικών επιπέδων ενάντια σε προϋπολογισμένα επίπεδα.

Εντέλει ο συνδυασμός αυτών των συστατικών προβλέψεων, δηλαδή των αποχωρήσεων των υπαρχόντων πελατών και των προβλέψεων αποχωρήσεων για τους νέους πελάτες, καθιστούν πιθανή την ανάπτυξη εκτιμήσεων των επιπέδων πελατών στο μέλλον.

3.5 Η Εξόρυξη δεδομένων στον κύκλο ζωής του πελάτη.

Για να είναι αποτελεσματικό ένα CRM σύστημα, θα πρέπει να συνδεθούν τα προϊόντα και οι στρατηγικές της εταιρείας με τους στόχους και τους πελάτες της. Ο όρος «κύκλος ζωής του πελάτη» (costumer life cycle) αναφέρεται σε όλα τα στάδια της σχέσης μεταξύ της επιχείρησης και του πελάτη. Είναι απολύτως απαραίτητη η κατανόησή του όρου από την εταιρεία, διότι σχετίζεται άμεσα με την κερδοφορία της (Rygielski. 2002). Σύμφωνα με τον Edelstein (2000), ο κύκλος ζωής του πελάτη αποτελείται από τρία στάδια, τα οποία είναι : η απόκτηση του πελάτη, η αύξηση της αξίας του πελάτη και η διατήρησή του.

1.Απόκτηση πελατών: αποτελεί το πρώτο βήμα του CRM και η εξόρυξη δεδομένων μπορεί να συνεισφέρει στη βελτίωση της αποτελεσματικότητας μιας εκστρατείας για την απόκτηση πελατών και στην ελαχιστοποίηση του κόστους.

2.Αύξηση της αξίας των υφιστάμενων πελατών:

- **Σταυροειδείς πωλήσεις**. Η εξόρυξη δεδομένων, χρησιμοποιώντας τις πληροφορίες των πελατών στις βάσεις δεδομένων, είναι σε θέση να βοηθήσει τον εκπρόσωπο εξυπηρέτησης πελατών να προτείνει τα κατάλληλα επιπρόσθετα προϊόντα στον πελάτη, ή και να μην προτείνει κανένα προϊόν, εάν ο πελάτης δεν είναι δεκτικός σε τέτοιου είδους πωλήσεις. Επίσης μέσω της εξόρυξης δεδομένων μπορούν να ελαχιστοποιηθούν τα παράπονα πελατών και να αυξηθεί η κερδοφορία της επιχείρησης.
- **Εξατομίκευση πελατών.** Μέσω της συσταδοποίησης της εξόρυξης δεδομένων γίνεται εφικτή η ομαδοποίηση των παρεμφερών προϊόντων, έτσι ώστε κάθε φορά που κάποιος πελάτης δείχνει ενδιαφέρον για ένα προϊόν, η εταιρεία να προβεί στις συστάσεις για αγορά περισσότερων προϊόντων. Με βάση το προφίλ του πελάτη εντοπίζονται οι πελάτες που μπορεί να ενδιαφέρονται για νέα προϊόντα που προστίθενται στον κατάλογο.

3. Διατήρηση των κερδοφόρων: Για σχεδόν κάθε εταιρεία, το κόστος απόκτησης ενός νέου πελάτη υπερβαίνει το κόστος διατήρησης κερδοφόρων πελατών. Με την δημιουργία του προφίλ των επικερδών, αλλά και των μη επικερδών πελατών, καθίσταται εφικτότερη η διατήρησή τους ως πελάτες, και ο εντοπισμός των πελατών που δεν αποφέρουν σημαντικά έσοδα στην επιχείρηση, αλλά θα μπορούσαν να αποφέρουν στο μέλλον.

Ο κύκλος ζωής του πελάτη αποτελεί ένα καλό πλαίσιο για την εφαρμογή της εξόρυξης δεδομένων στο CRM, αφού είναι σε θέση να προβλέψει την κερδοφορία των δυνητικών πελατών, καθώς αυτοί γίνονται ενεργοί, πόσο καιρό θα είναι ενεργοί πελάτες και ποια είναι η πιθανότητα να παύσουν να είναι πελάτες. Βέβαια δεν θα είναι ακριβής προγνωστικός δείκτης για το πότε συμβαίνουν οι περισσότερες εκδηλώσεις του κύκλου ζωής, αλλά θα μπορεί να βοηθήσει την επιχείρηση να αναγνωρίζει πρότυπα στα δεδομένα των πελατών της, τα οποία είναι προβλέψιμα. Για παράδειγμα, θα μπορούσε να προβλέψει τη συμπεριφορά που περιβάλλει ένα ιδιαίτερο γεγονός του κύκλου ζωής (π.χ. συνταξιοδότηση) και να βρει άλλους πελάτες σε παρόμοια στάδια ζωής και κατ’ επέκταση, ανάλογες συμπεριφορές.

Κεφάλαιο 4

Συμπεράσματα-Προτάσεις

4.1 Συμπεράσματα

Με την αυτόματη ανακάλυψη γνώσης μέσα από τις βάσεις δεδομένων, η εξόρυξη δεδομένων χρησιμοποιεί εξελιγμένες στατιστικές αναλύσεις και τεχνικές δημιουργίας μοντέλων για να αποκαλύψει κρυμμένα πρότυπα, κανόνες και σχέσεις σε οργανωμένες βάσεις δεδομένων. Κατά την διάρκεια των τελευταίων 40 ετών τα εργαλεία και οι τεχνικές επεξεργασίας δομημένων πληροφοριών έχουν συνεχίσει να εξελίσσονται από τις βάσεις δεδομένων στις αποθήκες δεδομένων. Οι εφαρμογές στις αποθήκες δεδομένων έχουν καταστεί κρίσιμες για τις επιχειρηματικές δραστηριότητες.

Παρότι η εξόρυξη δεδομένων βρίσκεται ακόμα σε αρχικό στάδιο, ένα ευρύ φάσμα εταιριών και βιομηχανιών συμπεριλαμβανομένων και των εταιριών λιανικής πώλησης και των τραπεζών και των τηλεπικοινωνιών, χρησιμοποιούν ήδη εργαλεία και τεχνικές εξόρυξης δεδομένων για να επωφεληθούν από τα ιστορικά τους στοιχεία . Με τη χρήση των τεχνολογιών αναγνώρισης προτύπων αλλά και στατιστικές και μαθηματικές τεχνικές, η εξόρυξη δεδομένων βοηθάει τους αναλυτές να εντοπίζουν σημαντικά γεγονότα, σχέσεις, πρότυπα, εξαιρέσεις αλλά και ανωμαλίες, οι οποίες θα μπορούσαν να περάσουν απαρατήρητες. Επίσης ιδιαίτερα χρήσιμη είναι η εφαρμογή τους και στις αποφάσεις μάρκετινγκ σε θέματα όπως η διαχείριση πελατειακών σχέσεων, το διαδραστικό μάρκετινγκ σε πραγματικό χρόνο και την δημιουργία προφίλ πελατών. Για αυτό τον λόγο υπάρχει η ανάγκη για βαθύτερη κατανόηση της χρήσης της εξόρυξης δεδομένων και διαχείρισης της γνώσης για την υποστήριξη των αποφάσεων μάρκετινγκ.

Η εξόρυξη δεδομένων έχει αρχίσει να εξελίσσεται σε μεγάλο βαθμό για διάφορους λόγους: οι οργανισμοί συγκεντρώνουν περισσότερες πληροφορίες για της επιχειρηματικές τους δραστηριότητες, το κόστος της αποθήκευσης δεδομένων έχει μειωθεί δραστικά και φυσικά οι ανταγωνιστικές πιέσεις των επιχειρήσεων έχουν αυξηθεί. Άλλοι παράγοντες είναι η εμφάνιση πιέσεων για τον έλεγχο των υφιστάμενων επενδύσεων, αλλά και η ένδειξη του δείκτη κόστους/απόδοσης των συστημάτων των ηλεκτρονικών υπολογιστών. Οι εφαρμογές της εξόρυξης δεδομένων είναι πολλές: από την ανάλυση και τη πρόβλεψη της συμπεριφοράς

των πελατών, μέχρι και τον σχεδιασμό στοχευμένων εκστρατειών μάρκετινγκ. Τέτοιες εφαρμογές είναι στενά συνδεδεμένες με την βελτιστοποίηση του κύκλου ζωής των πελατών και την αποτελεσματική διαχείριση πελατειακών σχέσεων.

Η διαχείριση πελατειακών σχέσεων από τις επιχειρήσεις κρίνεται απαραίτητη, προκειμένου να παραμείνουν ανταγωνιστικές στη σημερινή αγορά. Η αποτελεσματική χρήση των πληροφοριών των πελατών συνεπάγεται την άμεση ανταπόκριση στις ανάγκες τους, και κατά συνέπεια στην κερδοφορία. Προϋπόθεση για μια επιτυχημένη επιχείρηση είναι η κατανόηση των πελατών και των αναγκών τους, και η εξόρυξη δεδομένων αποτελεί το μέσο-κλειδί.

Υπάρχουν έξι βασικές εργασίες που εκτελούνται με την εξόρυξη δεδομένων:

- ταξινόμηση
- εκτίμηση
- πρόβλεψη
- ομαδοποίηση συγγένειας
- συγκέντρωση
- περιγραφή και σκιαγράφηση

Η εξόρυξη δεδομένων χωρίζετε σε κατευθυνόμενη και σε μη κατευθυνόμενη. Τα πρώτα τρία είναι παραδείγματα της κατευθυνόμενης εξόρυξης δεδομένων. Αντίθετα η ομαδοποίηση συγγένειας όπως επίσης και η συγκέντρωση είναι παραδείγματα μη κατευθυνόμενης εξόρυξης δεδομένων. Τέλος η σκιαγράφηση είναι ένας περιγραφικός στόχος που μπορεί να είναι είτε κατευθυνόμενος είτε μη κατευθυνόμενος.

Οι σημαντικότερες τεχνικές εξόρυξης δεδομένων όπως έχουν περιγραφεί παραπάνω είναι τα δέντρα απόφασης, οι νευρωνικοί αλγόριθμοι, η προσέγγιση του κοντινότερου γείτονα, η ανάλυση συνδέσεων, η συσταδοποίηση, και οι γενετικοί αλγόριθμοι.

Η εξόρυξη δεδομένων αποτελεί ένα πολύ ισχυρό εργαλείο που θα πρέπει να χρησιμοποιείται με μέγιστη προσοχή για την αύξηση της ικανοποίησης των πελατών και την παροχή ποιοτικότερων και ασφαλέστερων προϊόντων. Δε θα ήταν υπερβολικά αισιόδοξο αν λέγαμε ότι η εξόρυξη δεδομένων έχει ένα πολλά υποσχόμενο μέλλον και ότι τα επόμενα χρόνια θα φέρουν πολλές νέες εξελίξεις, μεθόδους και τεχνολογικές καινοτομίες. Επιπλέον, η βελτιωμένη ενοποίηση των τεχνικών εξόρυξης δεδομένων μπορεί να επιφέρει τη χρησιμοποίηση νέων τύπων δεδομένων και εφαρμογών.

4.2 Προτάσεις

Στα πλαίσια της παρούσας εργασίας κατέστη δυνατό να διερευνηθούν οι βασικές τεχνικές εξόρυξης δεδομένων καθώς επίσης και οι εφαρμογές των εργασιών της εξόρυξης δεδομένων στον κλάδο των επιχειρήσεων όπως για παράδειγμα των τραπεζικό τομέα η της εφαρμογές της στο λιανικό εμπόριο. Κατά συνέπεια, θα πρέπει να πραγματοποιηθεί περαιτέρω έρευνα για την εφαρμογή των τεχνικών εξόρυξης δεδομένων σε περισσότερους κλάδους

Σε βιβλιογραφικό επίπεδο, λόγω του περιορισμένου μεγέθους της συγκεκριμένης εργασίας, η μελέτη περιορίζεται στον αριθμό των ακαδημαϊκών πηγών, σε σχέση με τις έννοιες των όρων, τις μεθόδους και τις τεχνικές που αναφέρονται στην παρούσα μελέτη. Για τον λόγο αυτό, επιπρόσθετη έρευνα κρίνεται αναγκαία , προκειμένου να επιτευχθεί η σε βάθος κατανόηση όλων των πτυχών της εργασίας.

Η εξόρυξη δεδομένων, ως ένας σχετικά νέος και πολλά υποσχόμενος τομέας, εξακολουθεί να αντιμετωπίζει πολλές προκλήσεις και ανεπίλυτα προβλήματα που θέτουν νέα ερευνητικά ζητήματα για περαιτέρω μελέτη. Τέτοια θέματα αποτελούν η ανάπτυξη μιας ευέλικτης και κατανοητής γλώσσας απόδοσης των αποτελεσμάτων, η ανάπτυξη τεχνικών εξόρυξης δεδομένων σε προηγμένα συστήματα βάσεων δεδομένων, όπως οι ενεργές βάσεις δεδομένων (active databases) και οι χωρικές βάσεις δεδομένων (spatial databases) η ενσωμάτωση της ανακαλυφθείσας γνώσης με τη γνώση των εμπειρογνωμόνων, και η ανάπτυξη μεθόδων για τη διασφάλιση της ασφάλειας και την προστασία της ιδιωτικής ζωής.
Με την μετάβαση από το μαζικό μάρκετινγκ (mass marketing) στο εξατομικευμένο μάρκετινγκ (one-to-one marketing), και η περαιτέρω ανάπτυξη των εργαλείων εξόρυξης δεδομένων μπορεί να ωφελήσει όλες τις λειτουργίες του μάρκετινγκ. Η συστηματική εφαρμογή των τεχνικών εξόρυξης δεδομένων θα ενισχύσει τη διαδικασία της διαχείρισης γνώσης και θα εφοδιάσει την επιχείρηση με τις απαραίτητες γνώσεις για την αναγνώριση των αναγκών των πελατών και την καλύτερη εξυπηρέτησή τους. Η απόκτηση δεδομένων, για την βαθύτερη κατανόηση των πελατών, θα πρέπει να γίνεται με μη-παρεμβατικό τρόπο, με χαμηλό κόστος και υψηλή ακρίβεια. Επειδή η συμπεριφορά των πελατών είναι αβέβαιη και πολλές φορές, αντιφατική, οι ερευνητές θα πρέπει να κατασκευάσουν ένα δυναμικό μοντέλο τμηματοποίησης πελατών το οποίο θα αντανακλά αυτά τα χαρακτηριστικά.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ξένη:

Francis Buttle (2009) "Customer Relationship Management Concept and Technologies" ,Elsevier Ltd. σελ 93-102, 114-119,

Michael J.A. Berry, Gordon S. Linoff (2004) "Data Mining Techniques For Marketing, Sales, and Customer Relationship Management", Second Edition, Wiley Publishing, Inc. σελ.26-30, 43-86, 88-120

Tamanna Bhatia (2011) " Link Analysis Algorithms For Web Mining" ,International Journal of Computer Science and Technology, Vol. 2, Issue 2, June 2011, pp. 243-245.

Alex Berson, Stephen Smith, and Kurt Thearling (1999) "Building Data Mining Applications for CRM ". McGraw-Hill Companies

David L. Olson & Dursun Delen (2008) "Advanced Data Mining Techniques", Springer. σελ. 125-135

Paolo Giudici, Silvia Figini (2009) ''Applied Data Mining for business and industry'', Second Edition., John Wiley & Sons Ltd, σελ. 71-89

Jiawei Han, Micheline Kamber (2000) ''Data Mining Concepts and Techniques''. Morgan Kaufmann Publishers

Robert Elliot (2001) ''Data Mining cook book''. John Wiley & Sons, Inc, σελ. 25-26

Daniel T. Larose (2005) 'Discovering Knowledge in data- An Introduction to Data Mining , John Wiley & Sons, Inc σελ.11-17

Pascal Poncelet, Maguelonne Teisseire, Florent Masseglia (2008) ''Data Mining patterns new methods and applications'', Information Science Reference

Hussen Aly Abbass, Ruhul Amin Sarker, Charles S. Newton (2002) ‘‘Data mining a Heuristic Approach’’, Idea Group Publishing

Tsau Young Lin, Ying Yie, Anita Wasilewska, Churn-Jung Lian (Eds) (2008) ‘‘Data Mining : Foundations and Practice’’. Springer

Thanuja V., Venkateswaln B., and Anjanjeyulu G.S.S.N. (2011) ‘‘Applications of data mining in Customer Relationship Management’’, Journal of Computer and Mathematical Sciences, Vol 2

E.W.T. Ngai, Li Xiu, D.C.K. Chau (2008) “Application of data mining techniques in customer relationship management: A literature review and classification” Elsevier Ltd.

Parvatyar A. and Sheth J. N. (2001) Customer Relationship Management: Emerging Practice, Process and Discipline, Journal of Economic and Social Research Vol 3

Efstathios Kirkos, Charalambos Spathis, Yannis Manolopoulos (2007) “Data Mining techniques for the detection of fraudulent financial statements” Journal, Expert Systems with Applications: An International Journal. Volume 32 Issue 4, May, 2007 pp. 995-1003

Edelstein H. (2000) Building Profitable Customer Relationships with Data Mining, Executive briefing SPSS Inc , Two Crows Corporation.

Maimon, O. and Rokah, L. (2005). “Data Mining and Knowledge Discovery Handbook” Springer

Bose, R (2002) ‘‘Customer Relationship Management: Key components for IT success Industrial Management and Data Systems’’, Vol 102.

Konstantinos Tsiptsis, Antonios Chorianopoulos (2009) “ Data Mining Techniques in CRM - Inside Customer Segmentation” John Wiley & Sons, Ltd

Berry M. J. A. and Linoff G. S. (1999) ‘‘Mastering Data Mining The Art and Science of CRM’’, First Edition, New York, John Wiley and sons inc.

Rygielski C., Wang J. C. (2002) ‘‘Data Mining Techniques for Customer Relationship Management’’, Techonoly in Society Vol 24.

Forbes, September (2008)

Ελληνική:

Δημήτριος Β. Κοσμάτος (2011) “CRM Διαχείριση Πελατειακών Σχέσεων Αρχές και Τεχνολογίες” 2η Έκδοση, Κλειδάριθμος, σελ. 19

Παναγιώτης Α. Αντωνέλλης (2011) “Διδακτορική Διατριβή: Φιλτράρισμα και Εξόρυξη Δεδομένων σε Αντικείμενα Πληροφορίας”, Πανεπιστήμιο Πατρών, Πολυτεχνική Σχολή

Γεράσιμος Ε. Σταυλιώτης (2009), Εξόρυξη δεδομένων (Data Mining) Και Αναγνώριση Προτύπων Σε Κατηγορικά Δεδομένα Μέσω Συσταδοποίησης” Ελληνικό Στατιστικό Ινστιτούτο - Πρακτικά 22ου Πανελληνίου Συνεδρίου Στατιστικής (2009), σελ 201-210