Τ.Ε.Ι. ΠΕΛΟΠΟΝΝΗΣΟΥ

Σχολή Τεχνολογικών Εφαρμογών Σπάρτης

Τμήμα Μηχανικών Πληροφορικής Τ.Ε.

Data Mining και οι Εφαρμογές του στην Αγορά του Διαδικτύου

Μπρουμίδης Ευάγγελος

Επιβλέπων Καθηγητής: Γκατζιώλης Κλεάνθης

Σπάρτη, 2014

Ευχαριστίες

Θεωρώ υποχρέωση μου να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Γκατζιώλη Κλεάνθη για την βοήθεια που μου προσέφερε στην εκπόνηση αυτή της εργασίας, καθώς και τα στελέχη του ΤΕΙ Σπάρτης για τις πολύ εποικοδομητικές τους προτάσεις. Τέλος, ένα μεγάλο ευχαριστώ στην αρραβωνιαστικιά μου, Alex Hamper χωρίς την βοήθεια και την συμπαράσταση της οποίας δεν θα τα είχα καταφέρει.

Abstract

Η εργασία αυτή έχει ως στόχο την σύγκριση και μελέτη των σημαντικότερων τεχνικών και αλγορίθμων της επιστήμης της εξόρυξης γνώσης και τις κύριες εφαρμογές της στον σύγχρονο κόσμο του διαδικτύου. Στο πρώτο κεφάλαιο θα αναλυθούν οι κυριότεροι αλγόριθμοι που αποτελούν τον πυρήνα της επιστήμης καθώς και κάποιες βασικές προγραμματιστικές τους χρήσεις. Ακολούθως, το δεύτερο κεφάλαιο ασχολείται με τις τεχνικές συλλογής στοιχείων, τα οποία είναι η πεμπτουσία της εξόρυξης γνώσης μιας και προσφέρουν τον χώρο και την χρήση των τεχνικών της. Το τρίτο μέρος αφορά τους συνηθέστερους τόπους χρήσης των τεχνικών που προαναφέρθηκαν καθώς και τις ειδικές απαιτήσεις και προκλήσεις που θέτουν. Το τέταρτο κεφάλαιο αναλύει τις εφαρμογές της εξόρυξης γνώσης στην αγορά του διαδικτύου, δηλαδή τις κυριότερες χρήσεις της και τα προβλήματα που επιλύει. Τέλος, το πέμπτο κεφάλαιο περιλαμβάνει μία έρευνα σε μορφή ερωτηματολογίου που σκοπό έχει να βυθομετρήσει την γνώση του μέσου χρήστη σχετικά με την εξόρυξη γνώσης, την αγοραστική του φιλοσοφία και την γνώμη του σχετικά με την προστασία της ιδιωτικότητας. Η εργασία αυτή βοήθησε στην βαθύτερη κατανόηση της επιστήμης της εξόρυξης γνώσης και των εφαρμογών της σε κύριους τομείς της καθημερινότητάς μας.

This study aims to compare and analyze the most important techniques and algorithms of data mining and its most prominent applications in the modern internet world. In the first chapter, the main algorithms of the craft will be analyzed, as well as some of their basic programming iterations. Following this, the second chapter busies itself with documenting various data collection methods, which are the quintessence of data mining since they offer the ground basis for application of its ways. The third chapter mentions various places where data mining is used, as well as the specialized challenges it solves. The fourth chapter questions the deeper applications of data mining on the internet marketplace, as well as the problems it helped solve. Finally, the fifth and final chapter provides a survey in the form of a questionnaire in order to understand the average user's knowledge about the science of data mining, his buying

philosophy and his opinion on privacy protection. This study helped to attain a deeper understanding of the science of data mining and its applications in our everyday life.

Εισαγωγή

Τι είναι η εξόρυξη γνώσης; Είναι η ανάλυση ενός συνόλου δεδομένων και η εξαγωγή χρήσιμων συμπερασμάτων από αυτά. Σύνολα δεδομένων μπορούν να είναι ηλεκτρονικές βάσεις δεδομένων, στατιστικές αναλύσεις, και γενικότερα καθετί που περιέχει μεγάλες ποσότητες ταξινομημένων στοιχείων.

Σε αυτή την εργασία θα αναλυθούν οι κυριότεροι αλγόριθμοι που χρησιμοποιούνται από την επιστήμη της εξόρυξης γνώσης, καθώς και οι ειδικές χρήσεις τους. Έπειτα θα προβληθούν οι δημοφιλέστερες τεχνικές συλλογής στοιχείων προς επεξεργασία, τόσο οι κλασικές που προϋπήρχαν του διαδικτύου αλλά και μοντέρνες που κάνουν χρήση των σύγχρονων υπολογιστών και χρησιμοποιούνται από γίγαντες του διαδικτύου. Στη συνέχεια θα παρουσιαστούν οι κυριότεροι τόποι εφαρμογής των προηγούμενων τεχνικών, καθώς και οι ειδικές προκλήσεις που έθεσαν στην επιστήμη της εξόρυξης γνώσης και πως ξεπεράστηκαν. Το επόμενο κεφάλαιο αναλύει τις κυριότερες εφαρμογές της επιστήμης στην αγορά του διαδικτύου, αλλά και τα ισχυρότερα προγράμματα που εταιρείες ανάλυσης έχουν αναπτύξει. Τέλος, θα συζητηθούν τα αποτελέσματα ενός ερωτηματολογίου που τέθηκε τόσο χειρόγραφα όσο και στο διαδίκτυο. και αφορά την άποψη του μέσου χρήστη για την ιδιωτικότητα, τις διαφημίσεις και την εξόρυξη γνώσης γενικότερα.

Ο σκοπός αυτής της εργασίας είναι η βαθύτερη κατανόηση των εννοιών της εξόρυξης γνώσης και των τεχνικών της, αλλά και η έρευνα για να βρεθεί η κατάλληλη τεχνική για κάθε πρόβλημα. Για να επιτευχθεί αυτό έγινε έρευνα σε ένα πλήθος πηγών, τόσο βιβλίων όσο και διαδικτυακών ιστότοπων, αλλά και άρθρων και πτυχιακών. Λαμβάνοντας υπόψη το γεγονός ότι η εξόρυξη γνώσης ως όρος κατοχυρώθηκε στις αρχές του 1990, δεν παρουσιάστηκαν ιδιαίτερα προβλήματα στην ανεύρεση πηγών. Ας προχωρήσουμε στην βαθύτερη ανάλυση των κυριότερων αλγορίθμων εξόρυξης γνώσης.

Ιστορική Αναδρομή

Η εξόρυξη γνώσης ως αναγνωρισμένος όρος κατοχυρώθηκε από τα τέλη της δεκαετίας του 1990. η ίδια η επιστήμη όμως είναι πολύ αρχαιότερη. Πρωτόγονες μορφές στατιστικής και ανάλυσης των συμπερασμάτων που προκύπτουν υπάρχουν ήδη από την εποχή της αρχαίας Αιγύπτου. Η σύγχρονη όμως έννοια της εξόρυξης θεμελιώθηκε με την άνοδο των ηλεκτρονικών υπολογιστών και την δυνατότητα αποθήκευσης μεγάλων όγκων δεδομένων από το χαρτί σε δίσκους.

Η εξέλιξη των υπολογιστών από μικρής ισχύος και μεγάλου μεγέθους (μια σημερινή αριθμομηχανή τσέπης είναι ταχύτερη από έναν «υπερυπολογιστή» της δεκαετίας του 1960, ο οποίος είχε μέγεθος δωματίου) σε μικρά, ισχυρά και εύκολα προσβάσιμα εργαλεία ήταν το πρώτο βήμα της εξέλιξης. Το δεύτερο και σημαντικότερο ήταν η εισαγωγή των βάσεων δεδομένων και των γλωσσών προγραμματισμού που τις χρησιμοποιούν στις αρχές της δεκαετίας του 1980. Πλέον, τεράστιοι όγκοι δεδομένων μπορούσαν να αναλυθούν ταχύτατα και με ελάχιστη προσπάθεια από τον χρήστη. Η επανάσταση ήρθε όταν τα δεδομένα όχι μόνο μπορούσαν να περαστούν αυτόματα, αλλά και να απαντηθούν σύνθετα ερωτήματα όπως «τι ποσοστό πωλήσεων αναμένω τον επόμενο μήνα με βάση τους προηγούμενους;».

Αυτή η εξέλιξη οδήγησε την επιστήμη της εξόρυξης γνώσης να διασπαστεί σε τρεις τομείς: την στατιστική, την μηχανική μάθηση και την τεχνητή νοημοσύνη. Η μοντέρνα στατιστική κάνει χρήση των αυτοματοποιημένων μεθόδων που προσφέρουν οι υπολογιστές καθώς και την δύναμη του διαδικτύου συλλέγοντας πιο ποιοτικά δεδομένα. Η μηχανική μάθηση αφορά την εκπαίδευση αλγορίθμων στο να αναγνωρίζουν μοτίβα με την χρήση ενός σετ δοκιμαστικών δεδομένων, με ή χωρίς ανθρώπινη επίβλεψη. Τέλος, η τεχνητή νοημοσύνη που είναι η πλέον σύγχρονη τεχνολογία που έχει να επιδείξει η εξόρυξη γνώσης προσπαθεί να χρησιμοποιήσει ανθρώπινη λογική για να αναλύσει στατιστικές και να προβεί σε αποτελέσματα.

Ήδη από το 1989 υπάρχει παγκόσμιο συνέδριο για την εξόρυξη γνώσης και τις τεχνικές της, και από το 1997 έως και σήμερα υπάρχει το περιοδικό Data Mining and Knowledge Discovery το οποίο εκδίδει επιστημονικά άρθρα σχετικά με το αντικείμενο. Αν και σχετικά νέα επιστήμη (λιγότερο από 15 ετών) η εξόρυξη γνώσης συμπληρώνει την πολύ αρχαιότερη στατιστική με την δυνατότητα ανάλυσης των δεδομένων που εκείνη παρουσιάζει και με την εξαγωγή από αυτή

χρήσιμων συμπερασμάτων σε πολλούς σύγχρονους τομείς όπως το εμπόριο, το μάρκετινγκ, ο στρατός, η ιατρική και πολλούς άλλους.

Κεφάλαιο 1: Αλγόριθμοι Εξόρυξης Γνώσης

Εισαγωγή

Στο κεφάλαιο αυτό θα μελετηθούν οι δημοφιλέστεροι αλγόριθμοι που χρησιμοποιούνται για τη εξαγωγή συμπερασμάτων από βάσεις δεδομένων. Έχει δοθεί ιδιαίτερη βάση σε αυτούς που χειρίζονται ηλεκτρονικά δεδομένα και χρησιμοποιούνται κυρίως στο διαδίκτυο. Θα αναλυθούν οι κυριότερες λειτουργίες τους, η φιλοσοφία πίσω από την αρχιτεκτονική τους αλλά και η προσφορά τους στην επιστήμη της εξόρυξης γνώσης.

Από την κατασκευή του πρώτου αλγόριθμου με σκοπό την εξόρυξη γνώσης μέχρι τα σύγχρονα συστήματα που κάνουν χρήση τεχνητής νοημοσύνης υπήρχε η ανάγκη εξαγωγής αποτελεσμάτων με αυτοματοποιημένες διαδικασίες κάνοντας χρήση της ταχύτερης επεξεργαστικής ικανότητας των υπολογιστών και της νοητικής ικανότητας των ανθρώπων. Η κύρια χρήση τέτοιων συστημάτων είναι στην επιστήμη στης στατιστικής, αλλά και του μάρκετινγκ, κάνοντας χρήση σύνθετων αποτελεσμάτων που παρήγαγαν αυτοί οι αλγόριθμοι.

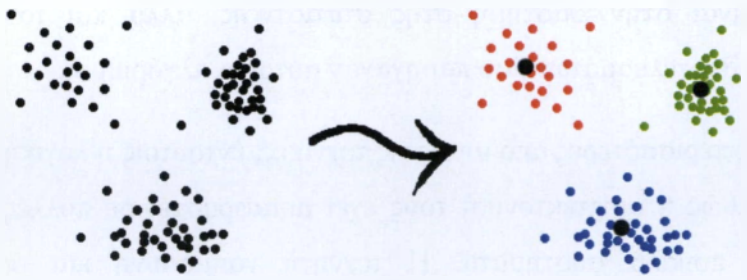
Αν και υπάρχει πηγαίος κώδικας για τις περισσότερες από αυτές τις τεχνικές, εντούτοις η λογική τους είναι αυτή που έχει σημασία, καθώς η αρχιτεκτονική τους έχει μεταφραστεί σε πολλές προγραμματιστικές γλώσσες και σε ποικίλα συστήματα. Η τεχνητή νοημοσύνη και οι τεχνολογικές εξελίξεις που ακολούθησαν βοήθησε αυτούς τους αλγορίθμους να εξελιχθούν στις σύγχρονες μορφές τους. Ας τους μελετήσουμε.

1.1 Clustering

Η τεχνική του clustering, ή ομαδοποίησης είναι μια κλασική τεχνική εξόρυξης γνώσης η οποία δημιουργεί στατιστικά μοντέλα βασιζόμενη σε στοιχεία με παρεμφερείς ιδιότητες, τα οποία ο αλγόριθμος διατάσσει σε ομάδες.

Η αρχή λειτουργίας του απαιτεί ένα πίνακα με σημεία σε τυχαίες θέσεις. Στην συνέχεια, επιλέγονται στοιχεία-κλειδιά είτε τυχαία είτε από τον χρήστη. Συνήθως αυτά τα στοιχεία είναι κοντά στον επιθυμητό μέσο όρο, αν και δεν είναι απαραίτητο να είναι απομακρυσμένα. Έπειτα μια διαδικασία ταυτοποίησης αποστάσεων μεταξύ τους (όπως τα Manhattan, Bregman και Ευκλείδεια) ανιχνεύει την απόσταση του σημείου κλειδιού από τα σημεία του πίνακα και αν βρεθεί κάποιο κοντινότερο τότε προστίθεται στην ομάδα (cluster) του πρώτου. Η διαδικασία συνεχίζεται έως ότου όλα τα στοιχεία έχουν ομαδοποιηθεί σε κάποιο σύνολο. Να σημειωθεί ότι κάθε τυχαίο στοιχείο συγκρίνεται με όλα τα σημεία κλειδιά για να αποφασιστεί σε ποιο cluster ανήκει.

Μία σημαντική λεπτομέρεια της ομαδοποίησης είναι ότι μπορεί να αναπαρασταθεί με τρόπους απεικόνισης κατανοητούς στον άνθρωπο σε κάθε στάδιο λειτουργίας του. Ο ίδιος ο αλγόριθμος μπορεί επίσης να εξελίξει τον εαυτό του θέτοντας νέα σημεία-κλειδιά σε περίπτωση που μεγάλος όγκος δεδομένων αλλάξει χαρακτηριστικά (για παράδειγμα η μέση ηλικία σε έναν πίνακα αυξάνεται από τα 18 στα 32). Ακόμα, η ομαδοποίηση είναι ένας γρήγορος τρόπος εξαγωγής συμπερασμάτων και μπορεί να αλλάξει αντικείμενο εστίασης αρκετά γρήγορα (να ομαδοποιήσει κατά ηλικία, έπειτα κατά φύλο, έπειτα κατά ύψος). Τέλος, μπορεί να προστατεύσει την βάση από διπλές εγγραφές, καθώς είναι εύκολο να ταυτολογηθούν, ειδικά σε μεγάλες βάσεις δεδομένων.



Γράφημα 1.1: Ο αλγόριθμος Clustering στην πράξη

Ένα μειονέκτημα της ομαδοποίησης ως τρόπο εξαγωγής συμπερασμάτων είναι το ότι ο αλγόριθμος πολλές φορές δεν μπορεί να αναγνωρίζει μοτίβα για πολύπλοκα δεδομένα χωρίς ανθρώπινη επιτήρηση να τον καθοδηγήσει. Επιπλέον, αν το σχήμα των δεδομένων πριν την

ομαδοποίηση δεν είναι σφαιρικό, δηλαδή να περιέχονται πολλά διαφορετικού βάρους δεδομένα τότε η αρχικοποίηση του αλγορίθμου ίσως να αποδώσει λανθασμένα στοιχεί κλειδιά. Επίσης, η ομαδοποίηση είναι σχετικά ευαίσθητη στον θόρυβο και τις εγγραφές εκτός ορίων (για παράδειγμα μια εγγραφή καταγράφει λόγω λάθους το ύψος ως 180 αντί 1,80) δημιουργώντας έτσι πρόβλημα στην αναζήτηση ενός μέσου όρου.

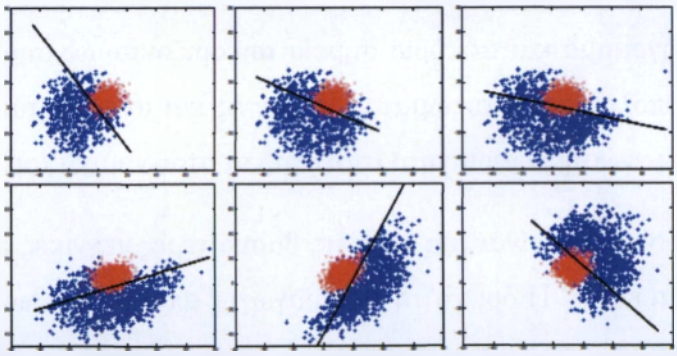
Παρ' όλα αυτά, η ομαδοποίηση έχει σημαντικές χρήσεις σε πολλούς τομείς, ο κυριότερος από τους οποίους είναι η στατιστική. Γνωρίζοντας την μέση τιμή του συνόλου των στοιχείων σε ένα γράφημα μπορούμε να ταξινομήσουμε γρήγορα νέες εγγραφές. Ακόμα, η χημεία χρησιμοποιεί ευρέως την ομαδοποίηση για να κατατάξει νέες χημικές ενώσεις ανάλογα με τις ιδιότητές τους και να διενεργήσει εικονικά πειράματα συνδυάζοντας ουσίες με παρεμφερείς ενέργειες και συντομεύοντας την διαδικασία πειραματισμού αποκλείοντας μη συνεργαζόμενες ενώσεις. Τέλος, αλγόριθμοι ομαδοποίησης χρησιμοποιούνται και από την επιστήμη της εγκληματολογίας με σκοπό την ταυτοποίηση υπόπτων. Η λογική εδώ είναι η αναζήτηση παρομοίων μοτίβων στα ήδη υπάρχοντα στοιχεία αντί για την ταξινόμηση νέων. Για παράδειγμα ένα αποτύπωμα εισάγεται στο πρόγραμμα και τα κύρια σημεία του ορίζονται ως σημεία-κλειδιά. Έπειτα αντιπαραβάλλεται με τα υπόλοιπα αποτυπώματα της βάσης και αναζητείται όμοιο. Αυτή η διαδικασία κάνει την εξόρυξη γνώσης απαραίτητο στοιχείο και στον νομικό τομέα.

Η ομαδοποίηση είναι μία από τις βασικότερες τεχνικές εξόρυξης γνώσης τόσο στο διαδίκτυο αλλά και εκτός. Η δυνατότητα παραγωγής αποτελεσμάτων τα οποία σχετίζονται μεταξύ τους με συγκεκριμένα χαρακτηριστικά αλλά και η δυνατότητα χειρισμού νέων εγγραφών με βάση τις παλιές είναι τα βασικότερα πλεονεκτήματά της. Σε κατάλληλα προετοιμασμένο περιβάλλον (με απουσία θορύβου ή λανθασμένων εγγραφών) η ομαδοποίηση είναι η αποδοτικότερη λύση για την εξαγωγή χρήσιμων συμπερασμάτων.

1.2 Adaboost

Ο αλγόριθμος Adaboost είναι μία τεχνική μηχανικής μάθησης που χρησιμοποιείται για την εξαγωγή συμπερασμάτων από ένα σύνολο εγγραφών σε μια βάση δεδομένων όταν υπάρχουν πολλοί κανόνες ανάγνωσης.

Ο λειτουργία του βασίζεται σε μία καινοτόμα διαδικασία, για την οποία οι δημιουργοί του Yoav Freund και Robert Schapire κέρδισαν το διεθνές βραβείο Godel το 2003. Το σύνολο των εγγραφών της βάσης χωρίζεται σε κομμάτια ν, τα οποία αναλύονται ξεχωριστά και δίνουν στατιστικούς ταξινομητές. Έπειτα αυτοί οι ταξινομητές σχηματίζουν ένα πρώτο σύνολο και δοκιμάζονται στην βάση δεδομένων, και τα αποτελέσματα που προκύπτουν αξιολογούν την βαρύτητα του κάθε κανόνα. Στην συνέχεια, κανόνες με χαμηλή βαρύτητα διαγράφονται και η διαδικασία ξεκινά ξανά έως ότου να υπάρχει ένας κανόνας που να εκφράζει το σύνολο των εγγραφών.



Γράφημα 1.2: Στάδια λειτουργίας του Adaboost

Για παράδειγμα, ας θεωρήσουμε ένα πίνακα Α με αριθμούς. Οι κανόνες που δίνονται είναι ότι δεν υπάρχουν αρνητικοί αριθμοί, ότι δεν υπάρχει το 0, ότι όλοι οι αριθμοί διαιρούνται με το 2 και ότι ο αριθμός 2 υπάρχει τουλάχιστον 10 φορές. Έπειτα από ένα πέρασμα σε υποσύνολο της βάσης διαπιστώνεται ότι όντως δεν υπάρχουν αρνητικοί αριθμοί (πρώτος κανόνας: βαρύτητα 2), υπάρχει ο αριθμός 0 (δεύτερος κανόνας: βαρύτητα 0), κάποιοι αριθμοί δεν διαιρούνται με το 2 (τρίτος κανόνας: βαρύτητα 0) και το 2 εμφανίζεται 12 φορές (τέταρτος κανόνας: βαρύτητα 2). Έστω ότι μετά από διαδοχικά περάσματα του συνόλου της βάσης οι κανόνες έχουν τις εξής βαρύτητες: πρώτος κανόνας(8), δεύτερος κανόνας(0), τρίτος κανόνας(0), τέταρτος κανόνας(7) από αυτό μπορούμε να εξάγουμε το συμπέρασμα ότι το σύνολο της βάσης δεν διαθέτει αρνητικούς αριθμούς, και ότι ο αριθμός 2 εμφανίζεται τουλάχιστον 12 φορές. Οι άλλοι δύο δεν επαληθεύτηκαν και άρα απομακρύνθηκαν.

Αυτή η λειτουργία του αλγόριθμου, το να αλλάζει δηλαδή τους κανόνες αντί για τα αποτελέσματα είναι ιδιαιτέρως χρήσιμη αν δεν γνωρίζουμε το πλήθος ή τον τύπο των στοιχείων μιας βάσης, ή αν θέλουμε να διατρέξουμε συγκεκριμένα ερωτήματα (υπάρχει η εγγραφή KXB1920; Αν υπάρχει, υπάρχουν και άλλες που ξεκινούν από KXB;) και να πάρουμε απάντηση. Η αξιολόγηση των ταξινομητών ανάλογα με την βαρύτητά τους βοηθά επίσης την απάντηση λογικών ερωτημάτων, δηλαδή εκείνων που απαντώνται με ναι ή όχι, αλλά και πιο σύνθετα όπως ερωτήματα SQL. Διασπώντας την βάση δεδομένων σε κομμάτια έχουμε ταχύτερη επεξεργασία των ερωτημάτων.

Η αρχιτεκτονική του αλγόριθμου έχει ένα σημαντικό μειονέκτημα, και αυτό είναι η ευαισθησία στον «θόρυβο» και δεδομένα με μεγαλύτερες του φυσιολογικού τιμές (για παράδειγμα δίνοντας 180 αντί για 1,80 στην τιμή του ύψους) αποκλείοντας έτσι κανόνες που θα μπορούσαν σε άλλη περίπτωση να είναι σωστοί. Ακόμα, έχουν καταγραφεί περιπτώσεις χρήσεις του Adaboost που επέστρεψαν λανθασμένα αποτελέσματα που προγραμματιστικά είναι σωστά αλλά δεν είναι αποδεκτά από την ανθρώπινη λογική. Αυτό οφείλεται πολλές φορές στο δοκιμαστικό σετ ταξινομητών, το οποίο δεν έχει σχέση με την βάση δεδομένων. Για παράδειγμα, ένα σετ που αφορά στοιχεία ποδοσφαιρικών αγώνων θα είχε λίγη σημασία αν χρησιμοποιούνταν από τον Adaboost για ανάλυση μιας βάσης δεδομένων φοιτητών. Ακόμα και πίνακες με λανθασμένες εγγραφές πάνω του 5% μπορούν να οδηγήσουν σε σετ κανόνων με πάνω από 50% μη αποδεκτά αποτελέσματα.

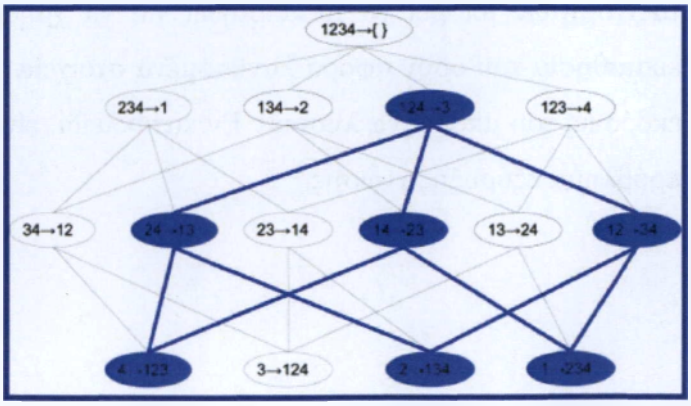
Ο αλγόριθμος Adaboost κατέχει μια σημαντική θέση ανάμεσα στις τεχνικές μηχανικής μάθησης και εξόρυξης γνώσης. Η λογική του είναι μοναδική, και εύκολα υλοποιήσιμη, αφού ένα σετ ταξινομητών μπορεί να μεταποιηθεί για να χρησιμοποιηθεί με πολλών τύπων δεδομένα. Η ευαισθησία του όσον αφορά λανθασμένα στοιχεία είναι ένα ελάττωμα, το οποίο οι μελλοντικές εκδόσεις του μπορεί να λύσουν. Εν κατακλείδι, είναι μια ισχυρή αριτεκτονική για σχεδόν κάθε πρόβλημα εξόρυξης γνώσης.

1.3 Apriori

Ο αλγόριθμος Apriori είναι μια από τις βασικότερες τεχνικές δημιουργίας μοτίβων και εξαγωγής συμπερασμάτων από στατιστικές. Η λειτουργία του βασίζεται στην ομαδοποίηση παρόμοιων αντικειμένων και την αναγνώριση της σχετικότητας νεοεμφανιζόμενων με την χρήση ενός κατώτατου ορίου ως μέσο σύγκρισης.

Για να κατανοήσουμε καλύτερα αυτή την αρχή, ας πάρουμε έναν υποθετικό πίνακα με προϊόντα από ένα σουπερ μάρκετ και ας θέσουμε τα ερωτήματα: από τους καταναλωτές που αγόρασαν καφέ, πόσοι αγόρασαν και ζάχαρη; Μπορούμε να υποθέσουμε με ασφάλεια ότι κάθε καταναλωτής αγοράζει καφέ μαζί με ζάχαρη; Υπάρχει σχέση ανάμεσα στον καφέ και το μέλι;. Ο αλγόριθμος ουσιαστικά διατρέχει τις ομάδες δεδομένων (καλάθια του καταναλωτή) προσπαθώντας να ανακαλύψει συσχετίσεις μεταξύ τους (καφές-ζάχαρη, πάνες-γάλα, μπύρα-πίτσα) αλλά και μη-συσχετιζόμενα αντικείμενα (σπάνια θα υπάρχει σύνδεση απορρυπαντικό-ξηροί καρποί). Κατ αρχάς, ανιχνεύει κάθε μοναδικό αντικείμενο και το τοποθετεί σε μια λίστα. Στην συνέχεια, αντικείμενα που εμφανίζονται λιγότερες φορές από έναν προκαθορισμένο μέσο όρο αφαιρούνται, αφήνοντας μόνο τα δημοφιλέστερα. Στην συνέχεα, δημιουργούνται σχέσεις ανάμεσα στα νέα αντικείμενα, και η διαδικασία επαναλαμβάνεται. Έτσι χτίζονται συνεχώς όλο και λιγότερες σχέσεις που είναι όμως όλο και δημοφιλέστερες έως ότου φτάσουμε σε μία και μοναδική σχέση που περιλαμβάνει όλα τα συχνότερα εμφανιζόμενα αντικείμενα.

Η χρήση του Apriori έφερε μεγάλες αλλαγές στην επιστήμη της εξόρυξης γνώσης. Πλέον υπήρχε η δυνατότητα εξαγωγής μοτίβων από τεράστιες βάσεις δεδομένων, αφού ο αλγόριθμος λειτουργεί αποδοτικότερα για μεγαλύτερο πλήθος στοιχείων. Η εφαρμογή του βρήκε μεγάλη απήχηση στο εμπόριο, εφόσον μεγάλα σετ από διαφορετικά αντικείμενα είναι και οι συναλλαγές. Ακόμη, η λογική του μπορούσε να χρησιμοποιηθεί και από λογισμικό παρακολούθησης πωλήσεων και στατιστικών όπως το GPower.



Γράφημα 1.3: Ο Apriori στην διαδικασία κατάταξης συνδέσμων. Οι ισχυρότεροι είναι τονισμένοι με μπλε.

Συνδυάζοντας την ταχύτατη υπολογιστική ισχύ των σύγχρονων υπολογιστών με την αποδοτικότητα του Apriori έχει καταστεί δυνατή η επεξεργασία σε πραγματικό χρόνο και με άμεσα αποτελέσματα υπό πραγματικές συνθήκες σε μεγάλη αλυσίδα σουπερ-μάρκετ της Αμερικής και τα αποτελέσματα ήταν άκρως ικανοποιητικά.

Ένα σημαντικό μειονέκτημα του Apriori είναι το ότι η διαδικασία συλλογής και ανακάλυψης μοτίβων είναι σχετικά χρονοβόρα, εφόσον απαιτούνται πολλά περάσματα από την βάση δεδομένων, η οποία μπορεί να είναι μεγάλη σε μέγεθος αλλά και να αλλάζει περιοδικά, καθιστώντας την διαδικασία αβέβαιη. Επιπλέον, υπάρχει πάντα η πιθανότητα να δημιουργηθούν πολλές μικρότερες σχέσεις χωρίς την δυνατότητα να ενοποιηθούν σε μεγαλύτερες. (για παράδειγμα πολλά σετ των τριών αντικειμένων, αλλά όχι με αρκετές συσχετίσεις μεταξύ τους για την δημιουργία σετ τεσσάρων αντικειμένων). Σε αυτή την περίπτωση ο αλγόριθμος τερματίζει με μη ικανοποιητικά αποτελέσματα. Τέλος, νεότερες εκδόσεις του Apriori έχουν την τάση να εξετάζουν μεγάλα κομμάτια μιας βάσης δεδομένων, αγνοώντας τα μικρότερα στην προσπάθειά τους για ταχύτερη έκδοση αποτελεσμάτων, χάνοντας πολλές φορές συσχετίσεις που περιλαμβάνονταν σε αυτά.

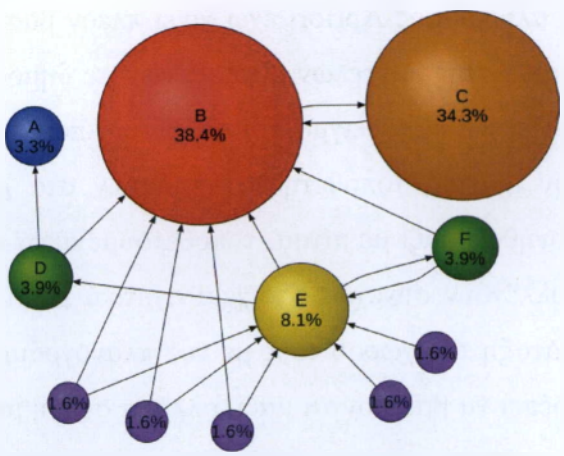
Ο αλγόριθμος Apriori έχει γίνει πλέον βασικό εργαλείο εξόρυξης γνώσης, αφού η δυνατότητα παραγωγής μοντέλων βασισμένων σε δημοφιλή στοιχεία μιας βάσης δεδομένων μπορεί να μας δείξει για παράδειγμα ποια προϊόντα πουλάνε περισσότερο αλλά και πιο σύνθετες έννοιες όπως την ιδανική τοποθέτηση προϊόντων στα ράφια (αν γνωρίζουμε ότι οι μπύρες αγοράζονται συνήθως μαζί με πίτσα, τοποθετούμε αυτά τα προϊόντα σε προσφορά μαζί). Τα σουπερ μάρκετ συλλέγουν συνεχώς στοιχεία σχετικά με τις προτιμήσεις των καταναλωτών και αλλάζουν την διάταξη του χώρου τους με νέα πλανογράμματα. Είναι προφανές ότι πάνες και βρεφικά γάλατα πρέπει να βρίσκονται μαζί, αλλά τι σχέση μπορούν να έχουν τα αλκοολούχα ποτά με τις πίτσες, τα αντικουνουπικά και τα φώτα ανοικτού χώρου; Εδώ εισέρχεται η ανθρώπινη νοημοσύνη αλλά και περιστασιακοί παράγοντες (στην παραπάνω περίπτωση μια αθλητική διοργάνωση το καλοκαίρι) για να εξάγει πραγματική γνώση από τα μοτίβα που προσφέρονται έτσι ώστε να παρθούν οι σωστότερες στρατηγικές αποφάσεις που θα αποφέρουν το μεγαλύτερο κέρδος.

1.4 PageRank

Ο αλγόριθμος PageRank είναι το κύριο εργαλείο που χρησιμοποιεί η Google για την κατάταξη των αποτελεσμάτων αναζήτησης, αναλύοντας και ταξινομώντας κάθε σελίδα ανάλογα με το πλήθος άλλων σελίδων που δείχνουν σε αυτή, καθώς και την ποιότητά τους. Κατασκευάστηκε από τους Sergey Brin και Larry Page το 1998.

Η λειτουργία του αλγορίθμου είναι η εξής: Οι λέξεις αναζήτησης που πληκτρολογεί ο χρήστης αναλύονται, και στην συνέχεα αναζητούνται σελίδες που ανταποκρίνονται στα κριτήρια. Για να αποδειχθεί ότι μια σελίδα είναι καταλληλότερη, ο PageRank καταμετρά πόσες άλλες «δείχνουν» σε αυτή και στην συνέχεια που «δείχνει» η ίδια. Η κάθε τέτοια ψήφος καθορίζει το πόσο σημαντική είναι. Δεν έχουν όλες οι ιστοσελίδες την ίδια βαρύτητα ψήφου όμως. Για παράδειγμα, ένα blog με θέμα τα είδη ζώων υπό εξαφάνιση έχει στους συνδέσμους του την κεντρική σελίδα της WWF, ενός παγκοσμίου οργανισμού. Κατά συνέπεια, δεν θα έχει μεγάλη σημασία για το ίδιο το WWF, εφόσον είναι ήδη μια πιστοποιημένη σελίδα. Αν όμως η WWF περιελάμβανε το blog στους καταλόγους της η θέση του θα ανέβαινε στην κατάταξη της Google.

Ένας σημαντικός παράγοντας που επηρεάζει την βαθμολογία μιας σελίδας είναι το πλήθος των εξερχομένων συνδέσμων. Αυτό σημαίνει ότι αν μια σελίδα «δείχνει» σε πολλές άλλες η βαρύτητά της μειώνεται. Έτσι εξασφαλίζεται η ακεραιότητα των γονέων-συνδέσμων και αποφεύγονται σελίδες-κατάλογοι, οι οποίες δεν εξυπηρετούν κανένα σκοπό. Ακόμα, το πόσο συχνά εμφανίζονται οι λέξεις κλειδιά και πως αυτές εκφράζονται έχει σημασία. Για παράδειγμα, μια σελίδα σχετική με φάρμακα θα πρέπει να βρίσκεται ψηλότερα από μια που απλώς αναγράφει την λέξη «φάρμακα» χιλιάδες φορές. Συνήθως, τέτοιου είδους ιστοσελίδες είναι αμφιβόλου ποιότητος.



Γράφημα 1.4: Ο αλγόριθμος PageRank. Οι μεγάλοι κύκλοι αντιπροσωπεύουν δημοφιλείς σελίδες.

Ο αλγόριθμος PageRank είναι η κύρια πηγή στατιστικών στοιχείων για την Google, λαμβάνοντας και ταξινομώντας καθημερινά εκατομμύρια σελίδες οι οποίες με την σειρά τους

δείχνουν σε άλλες. Η διαδικασία είναι πλήρως αυτοματοποιημένη και της επιτρέπει να γνωρίζει στατιστικά για κάθε ιστοσελίδα όπως η επισκεψιμότητα της, το πόσο ψηλά κατατάσσεται για αναζήτηση διαφόρων λέξεων κλειδιών, την αξιοπιστία της με βάση συγκεκριμένες σελίδες πιστοποίησης, και όλες αυτές οι πληροφορίες διατίθενται δωρεάν στους χρήστες μέσα από το Google Analytics. Η αναζήτηση αυτή μπορεί να γίνεται είτε μόνο στον τίτλο, είτε σε ολόκληρο το περιεχόμενο της σελίδας. Τα δεδομένα αυτά βοηθούν στην χαρτογράφηση του παγκόσμιου ιστού και κατευθύνουν τον χρήστη ανάλογα με τις προτιμήσεις του.

Τα πλεονεκτήματα του PageRank για τον χρήστη είναι η αξιοπιστία που δίνει μια υψηλή βαθμολογία στην σελίδα του, πράγμα που φέρνει και χρήματα από διαφημίσεις, αλλά και η προστασία από κακόβουλες ιστοσελίδες, οι οποίες φιλτράρονται. Μπορούμε επίσης να εξάγουμε γνώση μέσα από τις βάσεις δεδομένων της Google, τα αποτελέσματα της οποίας είναι αξιόπιστα, πολλές φορές ήδη σε μορφή έτοιμη για επεξεργασία.

Ένα σημαντικό μειονέκτημα του αλγορίθμου είναι το ότι νέες σελίδες με περιεχόμενο σχετικό με τις προτιμήσεις του χρήστη καθυστερούν να λάβουν υψηλή θέση εφόσον δεν υπάρχουν σελίδες που να «δείχνουν» σε αυτές. Ακόμα, είναι εύκολο να ανεβάσει κανείς την βαθμολογία του αγοράζοντας συνδέσμους σε μια σελίδα με ήδη ψηλό PageRank. Αυτή είναι μια συνηθισμένη τεχνική που χρησιμοποιούν σελίδες-αποθήκες συνδέσμων για να προβάλλουν τα περιεχόμενά τους (έναντι αμοιβής). Επιπλέον, ο αλγόριθμος κατατάσσει στοιχεία βασιζόμενος μόνο στις λέξεις, και όχι το νόημά τους. Για παράδειγμα ένας χρήστης που πληκτρολογεί την λέξη Steam, θα λάβει αποτελέσματα σχετικά με την γνωστή πλατφόρμα παιχνιδιών της Valve αντί για πληροφορίες για τον ατμό. Η ικανότητα να γίνεται αντιληπτό το νόημα αυτού που αναζητά ο χρήστης είναι το επόμενο βήμα στην εξέλιξη του PageRank και της εξόρυξης γνώσης γενικότερα.

Εν κατακλείδι, ο αλγόριθμος PageRank είναι ο πλέον σύγχρονος τρόπος κατάταξης ιστοσελίδων και εξαγωγής συμπερασμάτων για αυτές. Η οργάνωση αυτή βασίζεται στην απλή λογική του όσο περισσότεροι θεωρούν πιστοποιημένη μια σελίδα, τόσο ψηλότερα αυτή θα εμφανίζεται. Από το 1998 όταν πρωτοεμφανίστηκε μέχρι και σήμερα έχει επανασχεδιαστεί και συνεχώς βελτιώνεται για την καλύτερη δυνατή συλλογή, αξιολόγηση, οργάνωση και παρουσίαση αποτελεσμάτων. Το μέλλον θα δείξει αν θα υπάρξουν ανταγωνιστές, και πόσο επιτυχημένοι θα

είναι. Ένα είναι σίγουρο, η Google ρίσκαρε όταν τον χρησιμοποίησε και σήμερα είναι η πιο επιτυχημένη μηχανή αναζήτησης του παγκόσμιου ιστού.

1.5 Μπαεσιανή Λογική

Η Μπαεσιανή Λογική αφορά ένα σύνολο κανόνων πρόβλεψης της συμπεριφοράς μελλοντικών εισόδων με βάση κανόνες που ισχύουν για τις ήδη υπάρχουσες εγγραφές.

Για να γίνει πιο κατανοητή η λειτουργία του αλγορίθμου, ας παρατηρήσουμε το εξής σενάριο: Ένας φρουρός του αεροδρομίου πρέπει να αποφασίσει ποιοι επιβάτες μπορεί να είναι επικίνδυνοι για την ασφάλεια της πτήσης και να τους απομονώσει για επιπλέον έρευνα. Ξεκινώντας, δίνει σε κάθε χαρακτηριστικό των ανθρώπων μηδενική βαρύτητα για την απόφασή του (φύλο, ηλικία, εθνικότητα, αν κρατάει μεγάλη βαλίτσα, αν είναι νευρικός, αν χτύπησε ο ανιχνευτής μετάλλου και άλλους). Όσο προχωράει η μέρα, αρχίζει να αναγνωρίζει περιπτώσεις όπου ο επιβάτης όντως αποτελούσε απειλή (το 70% ήταν άντρες, οι γυναίκες δείχνουν πιο νευρικές, τα παιδιά δεν αποτελούσαν απειλή ακόμα και αν ο ανιχνευτής μετάλλων χτυπούσε, οι μεγάλες βαλίτσες κατά 30% είχαν κάτι μεμπτό). Έτσι, όταν ένας νέος επιβάτης φτάνει στο πόστο του, ο φρουρός παρατηρεί ότι είναι άνδρας, ιδιαίτερα νευρικός και σχετικά μεγάλος σε ηλικία. Βασιζόμενος στις παρατηρήσεις του έπειτα από εκατοντάδες επιβάτες μπορεί να αποφασίσει αν πρέπει να τον υποβάλει σε περαιτέρω έρευνα.

Ο αλγόριθμος μπορεί να συνθέσει κανόνες από γεγονότα που συνήθως συμβαίνουν μαζί (ένα μήλο είναι πάντα ένα φρούτο που είναι μικρό, στρογγυλό και κόκκινο), ή από γεγονότα που δεν συνδέονται ποτέ μεταξύ τους (ένα μήλο δεν είναι ποτέ φτιαγμένο από μέταλλο, οπότε καθετί μεταλλικό δεν μπορεί να είναι ποτέ φρούτο) αλλά και εν μέρει συνδεόμενα (μια ντομάτα είναι μικρή κόκκινη και στρογγυλή αλλά δεν είναι φρούτο). Η Μπαεσιανή λογική θεωρεί ότι κάθε κανόνας έχει την ίδια βαρύτητα ακόμα και αν είναι προφανείς (στο παράδειγμα του αεροδρομίου όλοι οι επιβάτες είναι άνθρωποι, δεν χρειάζεται να εξεταστεί αυτό) αλλά και το ότι όλοι οι κανόνες είναι ανεξάρτητοι μεταξύ τους, πράγμα που δεν αληθεύει στην πραγματική.

Όντως, το 2004 αποδείχθηκε ότι υπήρχαν σενάρια όπως έρευνες πάνω στον γενετικό κώδικα τα οποία περιέχουν ένα τόσο μεγάλο πλήθος κανόνων που τα κάνουν μη βιώσιμα σε περιβάλλον Μπαεσιανής λογικής. Βεβαίως, η ευκολία υλοποίησης του αλγόριθμου σε συνδυασμό με την δυνατότητα λειτουργίας του με ελάχιστα δεδομένα προπόνησης, εφόσον το σύστημα μαθαίνει όσο περισσότερο χρησιμοποιείται κάνουν την Μπαεσιανή λογική έναν ιδιαίτερα αποδοτικό τρόπο εξαγωγής συμπερασμάτων. Μία χαρακτηριστική χρήση όπου ο αλγόριθμος αποδίδει τα μέγιστα είναι σαν φίλτρο ανεπιθύμητης αλληλογραφίας, μαθαίνοντας λέξεις που πρέπει να αποκλείει από τις επιλογές του χρήστη.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Γράφημα 1.5 Μαθηματικός τύπος της Μπαεσιανής Λογικής.

Το πρόβλημα του αλγόριθμου είναι η μη έξυπνη προσέγγισή του σε ένα πρόβλημα. Όσο του προσδίδονται νέοι κανόνες θα συνεχίσει να τους αντιπαραβάλει με τα δεδομένα του ακόμα και αν δεν έχουν καμία λογική σχέση. Παραδείγματος χάριν, δοθέντος του κανόνα «ένα φρούτο είναι μαλακό» κάθε αντικείμενο της βάσης με την ετικέτα «μαλακό» μπορεί να θεωρηθεί φρούτο από τον αλγόριθμο ακόμη και αν οι υπόλοιπες ετικέτες του δεν περιγράφουν ένα σχετικό αντικείμενο.

Ολοκληρώνοντας, είναι σωστό να αναφερθεί ότι προγράμματα που κάνουν χρήση Μπαεσιανής λογικής έχουν κωδικοποιηθεί σε πολλές προγραμματιστικές γλώσσες όπως οι C#, Java, Python και Perl ενώ υπάρχει και μια έκδοση η οποία χρησιμοποιεί βασικές αρχές τεχνητής νοημοσύνης γραμμένη σε LISP, απαιτώντας έτσι ελάχιστη είσοδο από τον χρήστη. Η μελλοντική εξέλιξη του συστήματος ίσως φέρει βαθύτερη κατανόηση του νοήματος των ετικετών σε αντίθεση με την απλή ανάγνωση και την μηχανική εξέλιξη των κανόνων του συστήματος. Ένα είναι σίγουρο: ο αλγόριθμος της Μπαεσιανής λογικής είναι ένα εξαιρετικό εργαλείο εξόρυξης γνώσης από τεράστιους πίνακες δεδομένων και σίγουρα θα συνεχίσει να χρησιμοποιεί την αρχιτεκτονική του για πολύ καιρό ακόμα.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Κανόνας, Βαρύτητα, Φίλτρο, Μάθηση

Επίλογος

Σε αυτό το κεφάλαιο μελετήθηκαν ορισμένοι δημοφιλείς αλγόριθμοι, η αρχιτεκτονική των οποίων χρησιμοποιείται στα σύγχρονα προγράμματα εξόρυξης γνώσης. Από την ανάλυση προέκυψε ότι υπάρχουν πολλοί τρόποι να εξαχθούν συμπεράσματα από ένα σύνολο δεδομένων, και διαφορετικοί αλγόριθμοι ταιριάζουν σε διαφορετικές περιπτώσεις. Ακόμη έγινε αναφορά στα προβλήματα που έλυσαν, όπως το ζήτημα των στατιστικών ερευνών που διευκόλυνε ο Apriori και το ζήτημα της βέλτιστης διαδρομής το οποίο απάντησαν τα δεντροδιαγράμματα.

Ουσιαστικά, η αγορά του διαδικτύου έχει ανάγκη από ολοκληρωμένα συστήματα τα οποία θα λαμβάνουν ως στοιχεία εισόδου τεράστιους όγκους δεδομένων και θα πρέπει να είναι σε θέση να παράγουν συμπεράσματα σε σύντομο χρονικό διάστημα. Η επιστήμη της εξόρυξης γνώσης βρίσκεται σήμερα σε μια εξαιρετική φόρμα, χρησιμοποιώντας την μεγάλη επεξεργαστική ισχύ των σύγχρονων υπολογιστών και τους αλγορίθμους που αναφέρθηκαν σε αυτό το κεφάλαιο, καθώς και πολλούς άλλους.

Πώς όμως συλλέγονται τα δεδομένα που πρόκειται να επεξεργαστούν αυτοί οι ισχυροί αλγόριθμοι; Αυτό θα μελετηθεί στο επόμενο κεφάλαιο.

Κεφάλαιο 2: Τεχνικές Συλλογής Δεδομένων

Εισαγωγή

Σε αυτό το κεφάλαιο θα μελετηθούν οι δημοφιλέστερες τεχνικές συλλογής δεδομένων που χρησιμοποιεί η εξόρυξη γνώσης. Για την καλύτερη κατανόηση τους, είναι χωρισμένες σε δύο υποκατηγορίες: τις κλασικές, οι οποίες δημιουργήθηκαν και χρησιμοποιήθηκαν πριν την έλευση των υπολογιστών και τις σύγχρονες, οι οποίες κάνουν χρήση των νέων τεχνολογιών και του διαδικτύου και εφαρμόζονται κυρίως εκεί.

Φυσικά, υπάρχουν και άλλες τεχνικές πέρα από αυτές που θα αναφερθούν, οι οποίες όμως είτε χρησιμοποιούνται λιγότερο, είτε για ειδικούς σκοπούς (π.χ. στρατιωτικούς), είτε τέλος γιατί η πολυπλοκότητά τους τις καθιστά δύσχρηστες ή ακριβές σε σχέση με το αποτέλεσμα που παράγουν. Η σύγχρονη αγορά όμως χρειάζεται άμεσα και αξιόπιστα αποτελέσματα. Ας δούμε αναλυτικά τους τρόπους που τα παράγουν.

2.1.1. Στατιστική

Η αρχαιότερη και ταυτόχρονα η πιο σύγχρονη μέθοδος συλλογής στοιχείων και εξαγωγής συμπερασμάτων, η στατιστική υπήρχε ήδη από την αρχαία Ελλάδα, με τον πρώτο κατάλογο πλοίων για τον Τρωικό πόλεμο. Ως στατιστική ορίζεται ο κλάδος των μαθηματικών που συγκεντρώνει στοιχεία, τα ταξινομεί και τα παρουσιάζει σε κατάλληλη μορφή, ώστε να μπορούν να αναλυθούν και να ερμηνευτούν για την εξυπηρέτηση διαφόρων σκοπών.

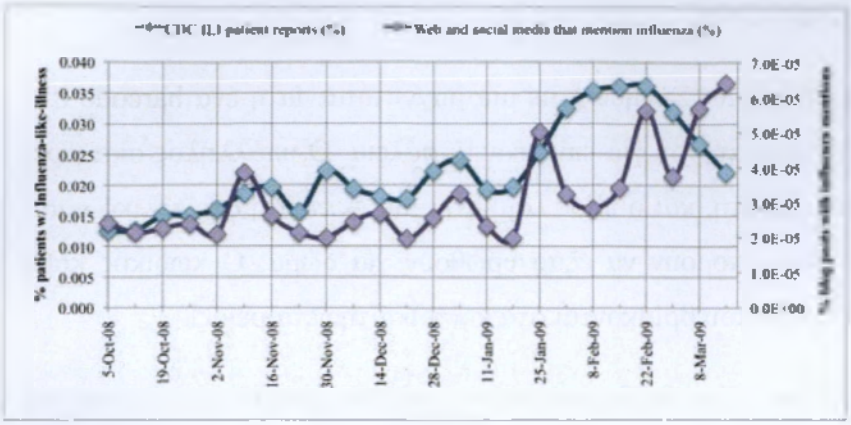
Η στατιστική λειτουργεί με διάφορους τρόπους ανάλογα με το θέμα που μελετείται. Για παράδειγμα, για να γνωρίσουμε τις τάσεις του εκλογικού σώματος θα χρησιμοποιήσουμε τηλεφωνικές δημοσκοπήσεις, ερωτηματολόγια, ή και ηλεκτρονικές δημοσκοπήσεις, τεχνικές που ανήκουν στην περιγραφική στατιστική. Αν θέλουμε τώρα να αναλύσουμε ένα πιο περιορισμένο σύστημα δεδομένων με αβέβαια αποτελέσματα, θα χρησιμοποιήσουμε τεχνικές αναλυτικής στατιστικής.

Υπάρχουν τέσσερα βασικά δομικά στοιχεία της στατιστικής. Η έννοια του λάθους, δηλαδή των αποτελεσμάτων που είναι εκτός των επιθυμητών ορίων ή δεν μπορούν να προβλεφτούν, η έννοια της ευστοχίας, δηλαδή το πόσο κοντά βρισκόμαστε στο πραγματικό ποσοστό που θέλουμε σε σχέση με αυτό που απεικονίζεται, η έννοια της ακρίβειας, ουσιαστικά το πόσο απέχει το κάθε στοιχείο από το άλλο, πόσο ομογενοποιημένη είναι η στατιστική μας, και τέλος η έννοια της προκατάληψης, το πόσο αποκλίνουν δηλαδή τα αποτελέσματα σε σχέση με τις προσδοκίες του ερευνητή.

Η χρήση της για την εξόρυξη γνώσης έχει να κάνει με την καταγραφή και ομαδοποίηση δεδομένων και είναι μια ιδιαίτερα δημοφιλής μέθοδος. Έχοντας ένα σύνολο καταγραφών, μπορούμε να τις ταξινομήσουμε και να τις ομαδοποιήσουμε, εξάγοντας σημαντικά συμπεράσματα. Για παράδειγμα, ένας πίνακας με στοιχεία φοιτητών όπως ονοματεπώνυμο, ηλικία, φύλο, και εξάμηνο σπουδών μπορεί να μας δώσει στοιχεία όπως την μέση ηλικία των φοιτητών, ποιο είναι το ποσοστό των Ελλήνων ή ακόμη και σύνθετα αποτελέσματα όπως το πόσα μαθήματα έχει περάσει ο καθένας βασιζόμενοι στο μέσο εξάμηνο. Αυτή η δυνατότητα εξαγωγής προχωρημένων αποτελεσμάτων από απλές συσχετίσεις είναι και η κύρια αξία της στατιστικής. Η εξόρυξη γνώσης βασίζεται στα μοτίβα που παράγονται και τα επεκτείνει.

Τα μειονεκτήματα της στατιστικής έγκειται περισσότερο στον τρόπο εισόδου των στοιχείων και τον ανθρώπινο παράγοντα. Παραδείγματος χάριν, ένας πίνακας μπορεί να παρουσιάζει σημαντική απόκλιση από το αναμενόμενο διότι το δείγμα που δόθηκε δεν ήταν αντιπροσωπευτικό, όπως και έγινε στις εκλογές του 1936 ανάμεσα τον Theodore Roosevelt και τον Alfred Landon. Η δημοσκόπηση της εφημερίδας Literary Digest έγινε σε μη αντιπροσωπευτικό κοινό (κυρίως υποστηρικτές του Landon), δίνοντας έτσι εσφαλμένες εκτιμήσεις για το αποτέλεσμα, που έφερε νίκη του Roosevelt με 62% έναντι 43% που γράφτηκε. Είναι σημαντικό το δείγμα να είναι πολύπλευρο και όσο το δυνατότερο διαφορετικό έτσι ώστε να υπάρχει σφαιρικότητα. Ο ανθρώπινος παράγοντας που επιφέρει λάθη είναι αφ' ενός οι προσδοκίες του αναλυτή, που συνήθως σημαίνει ότι το δείγμα είναι μεν σωστό αλλά αναμένονται διαφορετικά αποτελέσματα.

Το διαδίκτυο έφερε νέους τρόπους πραγματοποίησης στατιστικών και άρα εξόρυξης γνώσης. Πλέον οι ψηφοφορίες και οι δημοσκοπήσεις μπορούν να γίνουν διαδικτυακά, και μάλιστα με απεικόνιση αποτελεσμάτων σε πραγματικό χρόνο (strawpoll, surveymokey, easypolls). Ακόμη, δεδομένα εισρέουν και από ιστοσελίδες που πραγματοποιούν έρευνες ειδικού ενδιαφέροντος σε σχέση με το περιεχόμενό τους. Τέλος, στατιστικές οργανώνονται και με την χρήση ειδικών προγραμμάτων όπως το CSPRO, Που χρησιμοποιείται από την αμερικάνικη κυβέρνηση, τα ανοικτού κώδικα R και Shogun και το GNU για Linux.



Γράφημα 2.1.1: Στατιστική σε μορφή που μπορεί να παράγει συμπεράσματα. Εδώ το παράδειγμα μελέτης κρουσμάτων γρίπης.

Η στατιστική είναι ένας από τους καλύτερους τρόπους λήψης δεδομένων για εξόρυξη γνώσης, και από τους πιο κατανοητούς στο άνθρωπο. Η λογική της ομαδοποίησης και εύκολης παρουσίασης επιτρέπει την χρήση της στατιστικής σε μια ευρεία γκάμα σεναρίων, από εκλογικές αναμετρήσεις μέχρι χρηματοοικονομικές συναλλαγές και διαμόρφωση καταναλωτικών συμπεριφορών. Είναι βέβαιο ότι και στο μέλλον θα συνεχίσει έτσι, με ολοένα και αποδοτικότερους αλγόριθμους, για ολοένα και καλύτερα αποτελέσματα.

2.1.2 Κάρτες Μέλους

Ως εκπτωτική κάρτα ορίζεται η μέθοδος marketing η οποία χρησιμοποιει ένα συμβολικό μέσο (κάρτες, πόντους ή επιταγές) για την καταγραφή των αγοραστικών διαθέσεων των καταναλωτών με αντάλλαγμα εκπτώσεις ή δώρα.

Οι κάρτες σαν μέθοδος αναπτύχθηκαν στην δεκαετία του 80, με την καναδική εταιρεία Air Canada η οποία επιβράβευε τους συχνούς πελάτες με την χρήση ενός πρωτότυπου συστήματος

πόντων ανάλογα με τα διανυθέντα μίλια. Το σύστημα αυτό επεκτάθηκε και αναβαθμίστηκε με κύριο σημερινό χρήστη τα super market τα οποία ανταμείβουν τους πελάτες τους συλλέγοντας ταυτόχρονα τεράστιες ποσότητες δεδομένων.

Πως λειτουργεί όμως μια εκπτωτική κάρτα; Συνήθως, με μια μαγνητοταινία ή ένα barcode στο πίσω μέρος καθώς και τα στοιχεία του κατόχου για επιπλέον ασφάλεια. Ο υπάλληλος σκανάρει την κάρτα μαζί με τα προϊόντα του πελάτη, και ο ίδιος λαμβάνει μια έκπτωση στο τελικό ποσό πληρωμής ή ένα σύνολο πόντων που μπορούν να εξαργυρωθούν για δώρα. Ο κωδικός κάθε πελάτη είναι μοναδικός και τα δεδομένα του βρίσκονται στα κεντρικά της εταιρείας.

Το όφελος της επιχείρησης από την όλη διαδικασία είναι το ότι μπορεί να δημιουργεί αγοραστικά προφίλ για κάθε της πελάτη, συνθέτοντας τα δεδομένα από τις αγορές του για να γνωρίζει τις προτιμήσεις του, το βιοτικό του επίπεδο, ακόμη και αν έχει παιδία ή αν είναι χορτοφάγος. Η ίδια μέθοδος χρησιμοποιείται και για συνεργαζόμενες επιχειρήσεις. Για παράδειγμα, ποιές μάρκες καφέ είναι δημοφιλέστερες και ποιά μέθοδος διαφήμισης αποδίδει, με δυνατότητα απεικόνισης των στατιστικών αυτών.

Εικόνα 2.1.2: Διάφορες κάρτες μέλους και συλλογής πόντων.

Λόγω της απλότητας του τρόπου λειτουργίας τους, οι κάρτες επεκτείνονται πέρα από τα σουπερμάρκετ. Δίκτυα συνεργαζόμενων επιχειρήσεων οι οποίες αλληλοπροβάλονται και υποστηρίζουν την χρήση πόντων μεταξύ τους είναι πλέον μια κοινώς αποδεκτή τεχνικη μάρκετινγκ. Χαρακτηριστικό παράδειγμα το ελληνικό επιχειρηματικό πρόγραμμα "GO" της Εθνικής Τράπεζας, στο οποίο συμμετέχουν πάνω από 40 μεγάλες επιχειρήσεις κάθε κλάδου από αερογραμμές μέχρι κέντρα διασκέδασης και είδη σπιτιού.

Η εισαγωγή web υπηρεσιών έδωσε νέα ώθηση στις κάρτες και τα ανταποδοτικά προγράμματα γενικότερα. Πλέον, ακόμα και όταν βρίσκεται online, ο χρήστης μπορεί να λαμβάνει διαφήμιση σχετική με τις προτιμήσεις του που βρίσκονται ήδη αποθηκευμένες στον λογαριασμό του. ωΤέλος, υπάρχουν ήδη εφαρμογές για smartphones που καταργούν την ανάγκη για φυσικές κάρτες με ακριβώς τα ίδια προνόμια.

Όσο δημοφιλείς και να είναι οι κάρτες, δεν παύουν να υπάρχουν και οι αρνητές τους, με κύριο επιχείρημα την παραβίαση των προσωπικών τους δεδομένων και την καταγραφή των προτιμήσεών τους χωρίς την άδειά τους. Μερικά χρόνια πριν, ένα σκάνδαλο εμπορίας προσωπικών δεδομένων με σκοπό την διαφήμιση επέφερε πρόστιμα χιλιάδων ευρώ στους υπαίτιους καθώς και πολλά ερωτηματικά σχετικά με το πόσα πραγματικά στοιχεία καταγράφουν οι κάρτες μέλους. Ακόμα, κάποιοι πελάτες θεωρούν ότι οι συχνοί χρήστες των καρτών μέλους λαμβάνουν αδικαιολόγητα προνόμια σε σύγκριση με τους λιγότερο συχνούς, αποκλείοντας τους τελευταίους από ευκαιρίες που πιθανόν να τους ενδιέφεραν.

Είναι βέβαιο ότι οι κάρτες μέλους υπάρχουν και θα συνεχίσουν να υπάρχουν στην αγορά. Το πραγματικό ερώτημα είναι τι θα φέρει το μέλλον. Θα υπάρξει κάποτε μια ενιαία κάρτα για κάθε συναλλαγή; Και όσον αφορά την προστασία των προσωπικών δεδομένων ενάντια στις μεγαλύτερες και καλύτερες προσφορές, είναι δυνατόν να βρεθεί μια ισορροπία; Ήδη ο κόσμος εμπιστεύεται το σύστημα και οι επιχειρήσεις είναι ικανοποιημένες με τα δεδομένα που συλλέγουν. Είναι βέβαιο ότι πολλές ευκαιρίες για καινοτομίες θα εμφανιστούν.

2.1.3 Δέντρα αποφάσεων

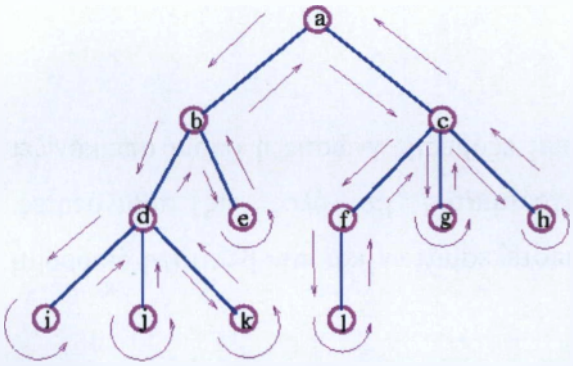
Δέντρα Αποφάσεων λέγεται μία τεχνική ανάλυσης και εξόρυξης γνώσης η οποία απεικονίζει κάθε δυνατό αποτέλεσμα σε μια μορφή διαγράμματος με όλες τις πιθανότητες. Χρησιμοποιούνται κυρίως για την μελέτη πιθανών αποτελεσμάτων και την βέλτιστη διαδρομή προς αυτά.

Σε ένα δέντρο αποφάσεων, κάθε «φύλλο» εκπροσωπεί ένα πιθανό τελικό αποτέλεσμα, ενώ κάθε κορμός μια η περισσότερες επιλογές που διακλαδώνονται έως ότου φτάσουν σε ένα τελικό στάδιο. Αυτή ακριβώς η λογική του να διακλαδώνονται μέχρι να μην υπάρχουν άλλοι συνδυασμοί κάνει τα δέντρα ιδανική μέθοδο για μηχανική μάθηση σε τεχνητή νοημοσύνη για παράδειγμα, αφού η μηχανή μπορεί να μελετήσει τον τρόπο σκέψης που προβάλλει το δέντρο και να τον κατανοήσει. Ήδη από την εφεύρεση του πρώτου δέντρου στα μέσα της δεκαετίας του 70, η κύρια χρήση τους ήταν η ανάλυση πολύπλοκων σεναρίων με πολλά πιθανά αποτελέσματα.

Η επιστήμη της εξόρυξης γνώσης χρησιμοποιεί τα δέντρα για την ανάλυση βάσεων δεδομένων και τον σχηματισμό διαγραμμάτων που δείχνουν τις συχνότερες τάσεις, παρόμοια δηλαδή με τις κλασικές στατιστικές τεχνικές. Για παράδειγμα, η απάντηση στο κλασικό ερώτημα της βέλτιστης διαδρομής ανάμεσα σε πέντε σημεία μπορεί να αναπαρασταθεί με ένα διάγραμμα δέντρου όπου κάθε κλάδος θα είναι ένα σημείο και κάθε «φύλλο» μια τελική λύση. Είναι προφανές ότι κάθε δυνατός συνδυασμός μπορεί να απεικονιστεί έτσι, και να παραχθούν συμπεράσματα για την ποιότητα κάθε αποτελέσματος.

Υπάρχουν κάποια προβλήματα όμως από την χρήση δέντρων για εξόρυξη γνώσης. Κατ αρχάς, υπάρχουν ιδέες που δύσκολα απεικονίζονται με ένα διάγραμμα, όπως πολύπλοκες μαθηματικές εξισώσεις, η λογικές πύλες, οι οποίες παράγουν αποτελέσματα που δύσκολα διαβάζονται από έναν άνθρωπο. Η συνηθέστερη λύση είναι η σύμπτυξη του προβλήματος είτε η διχοτόμησή του σε μικρότερα, και αναπαράσταση αυτών με ξεχωριστά δεντροδιαγράμματα για την εξαγωγή συμπερασμάτων. Ακόμα, για τις ανάγκες της μηχανικής μάθησης δημιουργούνται δέντρα τα οποία είναι ιδιαιτέρως πολύπλοκα μη δίνοντας έτσι ευκαιρία στο σύστημα να «μάθει» από απλά δεδομένα.

Τα μειονεκτήματα αυτά αντισταθμίζονται όμως από τις ιδιότητές τους. Η μεγάλη ευκολία που προσφέρουν τα δέντρα στην συγχώνευση στατιστικών, με την παρουσίασή τους σε μια πιο κατανοητή μορφή είναι ένα από τους λόγους που είναι ιδιαίτερα δημοφιλή. Επίσης, η δυνατότητα να χειρίζονται μεγάλα σετ δεδομένων, τα οποία δεν είναι απαραίτητο να είναι ταξινομημένα, ή ακόμα και σε κατάλληλη μορφή (για παράδειγμα αριθμητικές στατιστικές με ή χωρίς υποδιαστολή ή χαρακτήρες και αριθμούς ταυτόχρονα). Τέλος, ειδικότερα για εφαρμογές οικονομικού ενδιαφέροντος, ένα δεντροδιάγραμμα μπορεί να δείξει μοτίβα επενδύσεων που απέφεραν κέρδος και τι συνδυασμοί χρειάζεται να γίνουν για να καταλήξει κάποιος εκεί.



Γράφημα 2.1.3: Ένα δεντροδιάγραμμα. Το συγκεκριμένο επιτρέπει την κίνηση και προς τους επάνω κλάδους.

Είναι πλέον προφανές ότι τα δέντρα σαν τεχνική εξόρυξης γνώσης είναι ένας αποδοτικός τρόπος παρουσίασης και ταξινόμησης δεδομένων. Η απλή λογική τους σε συνδυασμό με την δυνατότητά τους να χειριστούν πολύπλοκους πίνακες δεδομένων τα κάνει ιδανικά για προγραμματιστικές εφαρμογές, την μηχανική μάθηση στα νευρωνικά δίκτυα αλλά και την χρήση τους ως μέσο ανάλυσης πινάκων και βάσεων δεδομένων. Το μέλλον ίσως φέρει δέντρα τα οποία να συνεργάζονται με την ολοένα και καλύτερη τεχνητή νοημοσύνη, χτίζοντας σενάρια και διαγράμματα που επιλύουν δύσκολα για την ανθρώπινη αντίληψη προβλήματα.

2.1.4 Νευρωνικά Δίκτυα

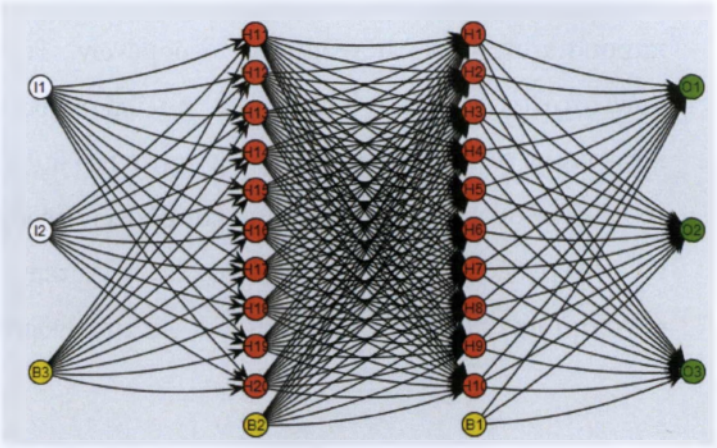
Ως νευρωνικό δίκτυο ορίζεται ένα σύνολο εργαλείων πρόβλεψης, ομαδοποίησης και παρουσίασης δεδομένων που βασίζεται στην έξυπνη παρακολούθηση και οργάνωση αυτών.

Το καλύτερο νευρωνικό δίκτυο είναι φυσικά ο ανθρώπινος εγκέφαλος, το πλεονέκτημα του οποίου ενάντια σε έναν υπολογιστή είναι η χρήση της λογικής. Ο υπολογιστής όμως αντισταθμίζει αυτό το μειονέκτημα με την ικανότητα εκτέλεσης πολύ περισσότερων εντολών σε ελάχιστο χρόνο, και άρα είναι το ιδανικό συμπλήρωμα του εγκεφάλου.

Η ιστορία των νευρωνικών δικτύων ξεκινά με την εισαγωγή των πρώτων υπολογιστών στην δεκαετία του 50 όπου με ειδικούς αισθητήρες πραγματοποιήθηκε το ακόλουθο πείραμα: μια σκούπα τοποθετήθηκε σε ένα ρομποτικό βραχίονα, ο οποίος υπολογίζοντας την κλίση της σκούπας προσπαθούσε να την κρατήσει σε ισορροπία. Παρά τα σημαντικά αποτελέσματα που προσέφερε το πείραμα στην μετέπειτα εξέλιξη της τεχνητής νοημοσύνης, δεν θεωρήθηκε επιτυχές. Κύριες αιτίες θεωρήθηκαν η χαμηλή υπολογιστική ισχύς των τότε μηχανημάτων καθώς και το υψηλό κόστος για την πραγματοποίηση περαιτέρω έρευνας.

Η επιστήμη των νευρωνικών δικτύων βρισκόταν σε τέλμα μέχρι το 1982, όταν ο John Hopfield εφηύρε την μηχανική μάθηση. Σε συνδυασμό με την απανθίζουσα τότε τεχνητή νοημοσύνη ήταν το εισιτήριο για την εξέλιξη των νευρωνικών δικτύων. Πλέον, ένα έξυπνο σύστημα μπορούσε να προβλέψει μελλοντικά δεδομένα μόνο διαβάζοντας παλαιότερες, επιβεβαιωμένες εγγραφές, και χτίζοντας πάνω στα μοτίβα που ανακαλύπτει.

Για να λειτουργήσει σωστά ένα νευρωνικό δίκτυο, χρειάζεται πρώτα ένα σετ δεδομένων για να «προπονήσει» το σύστημα. Το κλειδί για ένα επιτυχές νευρωνικό δίκτυο είναι το πλήθος των διαφορετικών δεδομένων του σετ, τα οποία πρέπει να περιέχουν αρκετούς τύπους εισόδου ώστε το σύστημα να είναι σε θέση να αναγνωρίσει και να διαχειριστεί κάθε μελλοντική είσοδο. Αν μετά από αρκετές δοκιμές το σύστημα αποδίδει ικανοποιητικά αποτελέσματα, είναι έτοιμα να δεχτεί πραγματικά δεδομένα.



Γράφημα 2.1.4 Αναπαράσταση ενός νευρωνικού δικτύου. Οι κόκκινες περιοχές είναι κανόνες που αλληλοσυγκρίνονται για να βρεθούν οι καταλληλότεροι.

Φυσικά, είναι αναμενόμενο όσο το σύστημα βρίσκεται σε κατάσταση μάθησης να έχει σαν έξοδο πολύ διαφορετικά δεδομένα εφόσον δεν έχει αναγνωρίσει ακόμη αρκετά μοτίβα. Εδώ θα παίξει σημαντικό ρόλο η ποιότητα της νοημοσύνης του συστήματος που θα του επιτρέψει να φιλτράρει, να ομαδοποιήσει και να αξιολογήσει την είσοδό του.

Οι σύγχρονες μέθοδοι εξόρυξης γνώσεως μπορούν να προσφέρουν ένα αξιόλογο πακέτο για επεξεργασία από ένα νευρωνικό δίκτυο, διευρύνοντας την ικανότητα επεξεργασίας τους. Η χρήση των νευρωνικών δικτύων για εξόρυξη γνώσης είναι σημαντικότατη, εφόσον η δυνατότητα της πρόβλεψης είναι η πεμπτουσία της στατιστικής. Συνδυάζοντας αυτά με τον «ανθρώπινο» τρόπο αντιμετώπισης των δεδομένων από την τεχνητή νοημοσύνη, δηλαδή λογική και μάθηση, έχει ως αποτέλεσμα μια έξοδο του προγράμματος κατανοητή και σωστή.

Η αύξηση της ταχύτητας και της επεξεργαστικής ισχύος των υπολογιστών καθιστά ικανή την επεξεργασία τεράστιων όγκων δεδομένων. Χαρακτηριστικό το παράδειγμα της AT&T η οποία κατάφερε την σύνδεση και συνεργασία 3 εκατομμυρίων νευρώνων σε 7 στρώματα με σκοπό την ανάλυση και πιστοποίηση επιταγών και διατραπεζικών συναλλαγών σε πραγματικό χρόνο με ασφάλεια και αξιοπιστία.

Όσο δημοφιλή και χρήσιμα και αν είναι τα νευρωνικά δίκτυα, δεν λείπουν και τα αδύνατα σημεία τους. Πρώτο και κύριο είναι το ότι μπορούν να μάθουν καλύτερα ανάλογα με το

δοκιμαστικό σετ δεδομένων που τους δίδεται. Ένα πολύ μεγάλο σετ μπορεί να κάνει το σύστημα αργό στην απόκριση, αφού πρέπει να ελέγξει και να συνδυάσει πολύ μεγαλύτερα σχήματα. Ένα μικρότερο σύνολο δεν θα έχει αρκετές περιπτώσεις για να μελετήσει το σύστημα και μπορεί να οδηγήσει σε σοβαρά λάθη. Το ίδιο το δείγμα πρέπει να βρίσκεται σε μορφή κατάλληλη για επεξεργασία (σωστά διαμορφωμένοι πίνακες, κατάλληλος τύπος εισόδου), και κυρίως να βρίσκεται σε όσο το δυνατόν μεγαλύτερη κατάσταση χάους έτσι ώστε να εξασκήσει το σύστημα στην αναγνώριση πολλών περιπτώσεων. Ιδανικό θα ήταν ένα ποσοστό 50-50 ανάμεσα σε σωστές και λάθος εγγραφές ώστε να μπορούν να ταυτολογηθούν και τα δύο είδη.

Τα νευρωνικά δίκτυα αποτελούν ένα αναπόσπαστο κομμάτι της εξόρυξης γνώσης, χρησιμοποιώντας τεχνητή νοημοσύνη και μηχανική μάθηση για να κάνουν υπολογισμούς με την λογική ενός ανθρώπου και την ταχύτητα ενός υπολογιστή. Η αποτελεσματικότητά τους τα καθιστά απαραίτητα σε εταιρείες στατιστικής, ενώ οι χρήσεις τους επεκτείνονται σε εκλογικές προβλέψεις, αθλητικά στοιχήματα και στρατιωτικές αναλύσεις με συνεχώς αυξανόμενες απαιτήσεις και διαρκώς εξελισσόμενη αρχιτεκτονική, η οποία παράγει όλο και καλύτερο λογισμικό.

2.2.1 Social Media Marketing

Με τη εμφάνιση των μέσων κοινωνικής δικτύωσης, μια νέα αγορά εμφανίστηκε για ένα νέο κοινό, το οποίο μοιραζόταν τις προτιμήσεις του ελεύθερα. Ως μέσο κοινωνικής δικτύωσης λέγεται ένα μέσο αλληλεπίδρασης ομάδων και ανθρώπων μέσω διαδικτυακών κοινοτήτων.

Στα πρώιμα στάδιά τους τα μέσα κοινωνικής δικτύωσης δεν είχαν ιδιαίτερη αξία για την επιστήμη της εξόρυξης γνώσης. Θεωρούνταν περισσότερο ένας τόπος ψηφιακής ψυχαγωγίας. Με την έλευση όμως του Web 2.0 απέκτησαν πιο δυναμικό περιεχόμενο, με περισσότερους χρήστες. Αυτό έφερε και νέους τρόπους συλλογής και αξιοποίησης γνώσης.

Πιο συγκεκριμένα το Facebook, αν και δημιουργήθηκε ως ένας τόπος επικοινωνίας συμφοιτητών, σήμερα είναι μια καθαρή αγορά, με μεγάλες επιχειρήσεις κάθε κλάδου να έχουν λογαριασμό και να διαφημίζουν τα προϊόντα και τις υπηρεσίες τους σε ένα μεγάλο πλήθος χρηστών, οι οποίοι μέσα από τους αλγόριθμους που χρησιμοποιούνται κατατάσσονται σε κατηγορίες ανάλογα με το πόσο πιθανόν είναι να ενδιαφέρονται για ένα συγκεκριμένο προϊόν. Αυτό επιτυγχάνεται αναλύοντας τις δηλωμένες προτιμήσεις του χρήστη και συνδέοντάς τες με σχετικές επιχειρήσεις. Με έναν απλό και δωρεάν λογαριασμό κάθε μικρή η μεγάλη επιχείρηση μπορεί να προσφέρει ειδικά προνόμια και ευκαιρίες στους χρήστες που την προβάλλουν (στην περίπτωση του Facebook με τα like). Αυτό το συμβιωτικό σχήμα ωφελεί και τους δύο.

Εικόνα 2.2.1: Τα βήματα προσέγγισης των κοινωνικών δικτύων από το μάρκετινγκ. Στρατηγική, Όνομα, Προφίλ, Διασυνδέσεις, Εκτέλεση.

Η χρήση μέσων κοινωνικής δικτύωσης αποδείχθηκε προσοδοφόρα και έτσι, ήδη από το 2001 πολλές επιχειρήσεις τα χρησιμοποίησαν για να προβληθούν. Με την χρήση κληρώσεων, κουπονιών, προσκλήσεων, ειδικών ευκαιριών με κωδικούς και άλλων τεχνικών μάρκετινγκ είναι δυνατή η αύξηση πωλήσεων με μηδενικό κόστος. Μια άλλη σχολή σκέψης θέλει επιχειρήσεις τύπου groupon να διαφημίζουν για λογαριασμό πολλών συνεργαζόμενων σε όσο το δυνατόν περισσότερους χρήστες ανεξαρτήτως προτιμήσεων. Οι θιασώτες της υποστηρίζουν ότι ακόμα και αν ένας χρήστης δεν έχει εκδηλώσει ενδιαφέρον, υπάρχει περίπτωση να αγοράσει αν παρουσιαστεί μια αρκετά καλή ευκαιρία.

Οι αντίθετοι της χρήσης των μέσων κοινωνικής δικτύωσης ως διαφημιστικό τοπίο δηλώνουν ότι η χρήση προσωπικών δεδομένων με σκοπό την διαφήμιση είναι παράνομο όπως παράνομη είναι και η παραβίαση της ιδιωτικότητας του ατόμου. Πράγματι, οι χρήστες βομβαρδίζονται από πλήθος διαφημίσεων από τις εταιρείες που παρακολουθούν, αλλά και από άλλες τις οποίες ουδέποτε είχαν σχέση. Επίσης, η εμπορεία προσωπικών δεδομένων από εταιρείες αμφιβόλου κύρους δεν είναι ανήκουστη. Το 2009 εξαρθρώθηκε εταιρεία από την Νιγηρία που πωλούσε σε

ενδιαφερόμενους «πακέτα» προσωπικών δεδομένων (όνομα, ενδιαφέροντα, λογαριασμό ηλεκτρονικού ταχυδρομείου, τηλέφωνο) προς 50 λεπτά του ευρώ το ένα.

Πρόβλημα όμως είναι μερικές φορές και ο ίδιος ο χρήστης, ο οποίος αποδέχεται όρους χωρίς να διαβάζει τα «ψιλά γράμματα», θέτοντας έτσι τον εαυτό του εκτεθειμένο ενάντια σε τέτοια άτομα. Μάλιστα από νομικής πλευράς εφόσον ο χρήστης αποδέχεται τους όρους θεωρείται ότι τους έχει πράγματι διαβάσει και συμφωνεί. Χαρακτηριστικό παράδειγμα ενός ηλεκτρονικού καταστήματος για υπολογιστές όπου στους όρους χρήσης υπήρχε ένα εδάφιο που ζητούσε από τον χρήστη να παραδώσει το πρωτότοκο παιδί του με όφελος σημαντικές εκπτώσεις. Μπορεί εδώ ο μόνος σκοπός να είναι το χιούμορ αλλά ο νόμος υποχρεώνει τις εταιρείες να αναγράφουν όλους τους όρους, αλλά όχι και τον χρήστη να τους διαβάζει.

Τα μέσα κοινωνικής δικτύωσης έχουν γίνει πλέον αναπόσπαστο κομμάτι του σύγχρονου διαδικτύου και οι αγορές δεν έμειναν αδρανείς, υιοθετώντας τα ως ένα μέσο δωρεάν διαφήμισης και ταυτόχρονης συλλογής δεδομένων προς επεξεργασία και βελτίωση των παρεχόμενων υπηρεσιών. Η εξόρυξη γνώσης βρίσκει πρόσφορο έδαφος, συλλέγοντας τις άμεσα διαθέσιμες πληροφορίες και ομαδοποιώντας τες. Όσο οι επιχειρήσεις χρησιμοποιούν αυτά τα δεδομένα με σεβασμό στην ιδιωτικότητα, η σωστή διαφήμιση βοηθά τον χρήστη να ανακαλύψει αυτό που ζητάει. Άλλωστε, η δικτύωση αυτή κάνει τον ευχαριστημένο πελάτη να φέρει και τον επόμενο.

2.2.2 Banners/Affiliates

Ο παλαιότερος τρόπος διαφήμισης, έχοντας αρχές από τα πρώτα στάδια του διαδικτύου ήταν το banner, ένα διαφημιστικό μήνυμα συνοδευόμενο από εικόνα και ήχο, στόχος του οποίου είναι να προσελκύσει ενδιαφερόμενους στην σελίδα που εκπροσωπεί. Οι affiliates είναι μια μετεξέλιξη των banner, όντας ουσιαστικά ιστότοποι με παρόμοιο περιεχόμενο τα οποία συνεργάζονται και αλληλοπροβάλονται έτσι ώστε να προσφέρουν την μεγαλύτερη ποικιλία επιλογών.

Ένα banner λειτουργεί με τον εξής τρόπο: Κατ' αρχάς crawlers και αλγόριθμοι εντοπισμού ανιχνεύουν το περιεχόμενο μιας σελίδας, συνήθως χρησιμοποιώντας λέξεις κλειδιά. Έπειτα ο

πελάτης μπορεί να αγοράσει διαφημιστικό χώρο ο οποίος εγγυάται ότι η διαφήμισή του θα προβάλλεται σε άτομα με σχετικά ενδιαφέροντα. Για παράδειγμα, διαφημίσεις για αυτοκίνητα δεν θα είχαν θέση σε μια σελίδα για παιδιά, και διαφημίσεις σχετικές με ταξίδια θα είχαν θέση σε έναν online ταξιδιωτικό οδηγό.

Το ίδιο το banner μπορεί να περιλαμβάνει οτιδήποτε ως μέσο προσέλκυσης, χωρίς αυτό να είναι πάντα θετικό. Αν και τα περισσότερα προβάλλουν κάτι σχετικό με τον ιστότοπο που τα φιλοξενεί πολλά χρησιμοποιούν παραπλανητικές τεχνικές όπως δυνατή μουσική, κρύψιμο πίσω από λειτουργίες της σελίδας ή ακόμη και να επανεμφανίζονται αφού ο χρήστης τα κλείσει. Ακόμη, η ανακατεύθυνση που κάνουν μπορεί να είναι σε ιστότοπους με κακόβουλο περιεχόμενο. Η διαγραφή τέτοιου είδους banners είναι δύσκολη καθώς η αυτοματοποιημένη διαδικασία που χρησιμοποιούν οι crawlers θεωρεί ότι το banner όντως έχει θέση σε μια σελίδα λόγω του πλήθους των σχετικών λέξεων κλειδιών ακόμη και αν η πραγματική λειτουργία του είναι κακόβουλη. Πρόσφατα, η Google έλαβε εντολή για την εκκαθάριση πολλών banner που διαφήμιζαν φάρμακα με επικίνδυνες παρενέργειες που ουδέποτε αναφέρθηκαν, κατασκευασμένα στην Κίνα τα οποία διαφημίζονταν σε κοινότητες σχετικές με την καταπολέμηση του καρκίνου.

Εικονα 2.2.2: Διάφορα banners που εμφανίζονται σε δημοφιλείς ιστοσελίδες.

Οι affiliates τώρα, αν και λειτουργούν με την ίδια αρχή, έχουν μια βασική διαφορά. Δεν διαδίδονται τυχαία μέσω αλγορίθμων, αλλά δημιουργούν δίκτυα μεταξύ τους, επιβεβαιώνοντας ο ένας την αξιοπιστία του άλλου. Για παράδειγμα, μια επιχείρηση εμπορίας ρούχων μπορεί να προσθέσει στο δίκτυό της μια άλλη που πουλάει παπούτσια και έτσι μέσα από την αλληλοδιαφήμιση να μεγιστοποιούν το κέρδος τους. Ο τρόπος που επιτυγχάνεται αυτό είναι με την χρήση affiliate links σε ένα σημείο της σελίδας τα οποία προτείνουν στον χρήστη προϊόντα

σχετικά με τις αγορές του. Σε συνέχιση του προηγούμενου παραδείγματος, όσο ο πελάτης αγοράζει ρούχα βλέπει και προσφορές για παπούτσια από τον συνεργαζόμενο ιστότοπο. Η όλη διαδικασία φυσικά αποφέρει ένα χρηματικό αντάλλαγμα στον ιδιοκτήτη ο οποίος προβάλει δωρεάν διαφήμιση για κάποιον άλλο.

Ουσιαστικά σκοπός αυτών των δικτύων είναι η δημιουργία ενός «κύκλου» που περικλείει πολλές σχετικές μεταξύ τους επιχειρήσεις, αλλά και σελίδες με νέα, αθλητικές, ιστότοποι ειδικού ενδιαφέροντος (ειδικά για προγραμματισμό και εφαρμογές) με σκοπό την κατάκτηση μεγαλύτερου αγοραστικού μεριδίου, η οποία θέλει άμεσα αποτελέσματα, πολλές επιλογές, και ποιότητα.

Η χρησιμότητα αυτών των δυο μεθόδων στην εξόρυξη γνώσης είναι σημαντική. Τα ίδια τα banners επιστρέφουν στατιστικές δείχνοντας τι ποσοστό επισκεπτών τα χρησιμοποίησαν, άρα δίνοντας και στατιστικά επισκεψιμότητας, και τα affiliate sites βοηθούν τους crawlers να εντοπίσουν περισσότερα αποτελέσματα για την ίδια αναζήτηση, προβάλλοντας τις λίστες των συνεργατών τους. Περισσότερα δεδομένα για εξόρυξη, καλύτερη χαρτογράφηση, ποιοτικότερα αποτελέσματα . Επιπλέον, μεγάλες εταιρείες στατιστικής αλλά και οι ίδιες οι σελίδες χρησιμοποιούν banners ως μικρούς στατιστικούς πίνακες ρωτώντας τους χρήστες σχετικά με τις προτιμήσεις τους και έτσι βελτιώνοντας την εμπειρία των χρηστών. Η στατιστική αυτή χρήση είναι δευτερεύουσα σε σχέση με την διαφήμιση, και πολλοί χρήστες απλώς αγνοούν κάθε διαφημιστική προσπάθεια διότι δεν θέλουν να δώσουν προσωπικά δεδομένα. Αυτό είναι φυσικά σεβαστό, και γι' αυτό υπάρχει νομοθεσία που υπαγορεύει την χρήση των προσωπικών δεδομένων που ένα banner μπορεί να ζητήσει και για πόσο καιρό αυτά θα αποθηκεύονται.

Πέρα από τα μειονεκτήματα των banners και των affiliates, τα οποία οφείλονται κυρίως στον τρόπο χρήσης και όχι στην αρχιτεκτονική τους, είναι η πλέον δημοφιλής μέθοδος διαφήμισης στον παγκόσμιο ιστό σήμερα. Η επιστήμη της εξόρυξης γνώσης χρησιμοποιεί και τα δύο για την άντληση δεδομένων και πληροφοριών με όλο και πιο αποτελεσματικές τεχνικές, κάνοντας την αγορά του διαδικτύου ένα καλύτερο μέρος για να βρει κανείς αυτό που ψάχνει, και πολλά ακόμα.

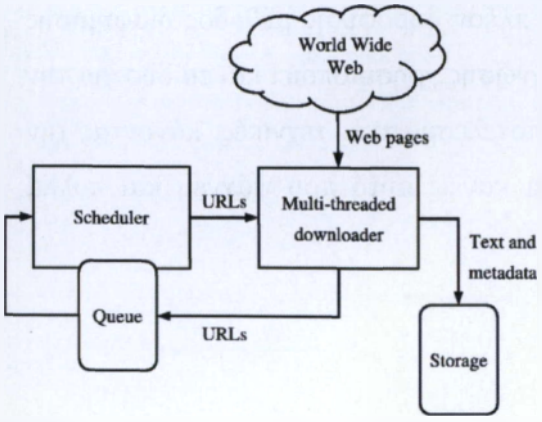
2.2.3 Web Crawlers

Ως ορισμός του web crawler θα ήταν ένα πρόγραμμα το οποίο επισκέπτεται το διαδίκτυο με μεθοδικό τρόπο με σκοπό την βελτιστοποίηση της διαδικασίας εύρεσης δεδομένων. Οι περισσότερες μηχανές αναζήτησης χρησιμοποιούν crawlers οι οποίοι αναλύουν συνεχώς νέους ιστότοπους και χαρτογραφούν το διαδίκτυο.

Η μέθοδος που χρησιμοποιούν βασίζεται στην ιδέα το ότι όσο περισσότερο ανταποκρίνεται ένας ιστότοπος στα κριτήρια αναζήτησης του χρήστη, τόσο ψηλότερα θα πρέπει να βρίσκεται στην λίστα αποτελεσμάτων. Για να επιτευχθεί αυτός ο σκοπός, ο crawler διαθέτει μια λίστα με ιστοσελίδες προς ανάλυση και αξιολόγηση. Επειδή η λίστα αυτή περιλαμβάνει πολλά εκατομμύρια site, ο crawler διαχειρίζεται πρώτα αυτά που έχουν αποδειχθεί αξιόπιστα, μειώνοντας έτσι τον φόρτο εργασίας του. Για κάθε διεύθυνση αναλύεται ο αριθμός άλλων διευθύνσεων που «δείχνουν» σε αυτή καθώς και αυτές στις οποίες «δείχνει» η ίδια. Έτσι καθορίζεται η βαρύτητα κάθε ιστότοπου, δηλαδή το πόσο σχετική είναι σε σύγκριση με συγκεκριμένες λέξεις κλειδιά.

Αυτή η διαδικασία είναι εξαιρετικά πολύπλοκη, αλλά τα επιστρεφόμενα αποτελέσματα σπάνια έχουν διακυμάνσεις η διαφορές. Αυτό οφείλεται στο γεγονός ότι αν ένας ιστότοπος πιστοποιηθεί ως αξιόπιστος σπάνια θα χάσει αυτήν την ιδιότητα. Για παράδειγμα, η σελίδα της Unicef χρησιμοποιεί ως λέξεις κλειδιά την ανθρωπιστική βοήθεια, την φιλανθρωπία και την φροντίδα ασθενών. Ένας τεράστιος αριθμός άλλων σχετικών σελίδων δείχνουν σε αυτή και έτσι ο crawler μπορεί να διαπιστώσει ότι έχει μεγαλύτερη βαρύτητα από ότι ένα blog για την μουσική.

Τα εργαλεία που χρησιμοποιούνται στην διαδικασία αξιολόγησης είναι τα εξής: Η δυνατότητα ομαδοποίησης παρόμοιων σελίδων για ταχύτερη επεξεργασία και άρα αποδοτικότερη χρήση του χρόνου. Ο χρόνος που ένας crawler διατρέχει μια ιστοσελίδα είναι ζωτικής σημασίας, μιας και αυξάνει κατακόρυφα την χρήση της κατά την διαδικασία της ανάλυσης, αποκλείοντας πιθανόν χρήστες που επιθυμούν να την επισκεφθούν.



Εικόνα 2.2.3: Απλοποιημένη αναπαράσταση του τρόπου λειτουργίας ενός crawler.

Για αυτό τον λόγο εφαρμόζεται μια τακτική γνωστή ως κώδικας «ευγένειας», δηλαδή μια πρώτη ανίχνευση της κίνησης την συγκεκριμένη ώρα και ένα χρονικό περιθώριο ανάμεσα σε κάθε ανάλυση προς αποφυγή υπερφόρτωσης των πόρων της σελίδας. Παρ' όλα αυτά, ορισμένες σελίδες χρησιμοποιούν κώδικα με οδηγίες προς τους crawlers με την ονομασία robots.txt όπου περιέχονται επιτρεπόμενες περιοχές ανάλυσης ή επιθυμία μη ανάλυσης. Το Robots.txt είναι γραμμένο σε html και εμφανίζεται στην αρχή μιας σελίδας. Αξίζει να σημειωθεί ότι ορισμένοι crawlers αγνοούν το robots.txt υιοθετώντας τον κώδικα ευγένειας αντί αυτού. Σε αυτές τις περιπτώσεις δημιουργείται σύγχυση και τα αποτελέσματα είναι συνήθως λανθασμένα λόγω του ότι ο crawler προσπαθεί να αναλύσει περιοχές στις οποίες δεν έχει πρόσβαση.

Η εξόρυξη γνώσης χρησιμοποιεί τα δεδομένα που επιστρέφουν οι crawlers μετά την ολοκλήρωση της ανίχνευσης μιας σελίδας. Μας ενδιαφέρει να γνωρίζουμε ποιοί είναι οι δημοφιλέστεροι ιστότοποι, τι είδους περιεχόμενο έχουν και τι κίνηση κατευθύνεται εκεί. Με αυτές τις πληροφορίες αποφασίζεται τι είδους διαφημίσεις παρέχονται έτσι ώστε να είναι σχετικές και αποτελεσματικές. Ακόμα, ανακαλύπτονται κενά στην αγορά από αναζητήσεις που δεν αποδίδουν ικανοποιητικά αποτελέσματα και ανακατευθύνονται σύνδεσμοι που βοηθούν στην αποσυμφόρηση των αναζητήσεων. Αυτό εφαρμόζεται κυρίως σε ιδέες και αντικείμενα που δεν προϋπήρχαν στο διαδίκτυο όπως μια φιλανθρωπική εκδήλωση ή μία καινούργια τεχνολογική εξέλιξη, γεγονότα τα οποία προσελκύουν κόσμο που ενδιαφέρεται να μάθει περισσότερα. Η δουλειά των crawlers είναι να παράγουν πληροφορίες ανιχνεύοντας συσχετίσεις από δημοφιλείς ιστότοπους.

Κλείνοντας, είναι σημαντικό να αναφερθούν κάποιοι από τους γνωστότερους crawlers . Ο γνωστότερος είναι ο Googlebot, της ομώνυμης εταιρείας, το Yahoo! Slurp αλλά και το Bingbot της Microsoft. Πέρα από τους γίγαντες όμως, υπάρχει και δωρεάν λογισμικό ανοικτού κώδικα για την εξυπηρέτηση μικρότερης κλίμακας ανίχνευσης, όπως το Dataspark, το Aspseek αλλά και το ειδικευμένο για προγραμματιστικές εφαρμογές 80legs, του πανεπιστημίου MIT το οποίο χρησιμοποιεί τους ελεύθερους πόρους πολλών υπολογιστών για να ανιχνεύσει ιστοσελίδες σε πραγματικό χρόνο.

2.4.4 Cookies

Ως cookies θα μπορούσαν να ονομαστούν μικρά πακέτα δεδομένων που στέλνονται από ιστοσελίδες προς τον φυλλομετρητή του χρήστη όσο εκείνος βρίσκεται στην συγκεκριμένη σελίδα και παραμένουν μετά το τέλος της περιήγησής του.

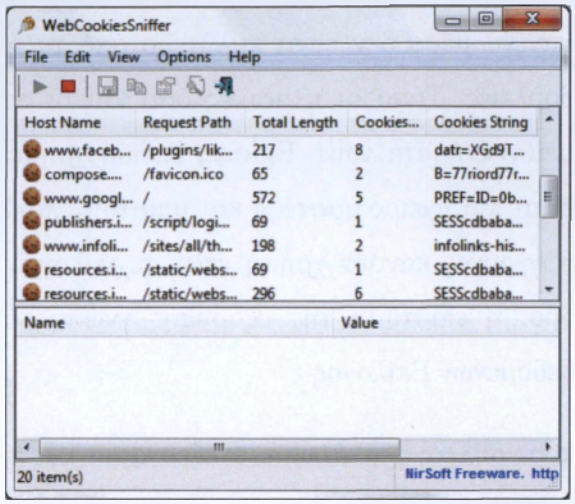
Ένα cookie είναι γραμμένο συνήθως σε JavaScript και περιέχει πληροφορίες όπως το όνομα του χρήστη, συγκεκριμένες επιλογές που έκανε ή αποθηκευμένους κωδικούς, με σκοπό να κάνει την περιήγησή του ευκολότερη την επόμενη φορά που θα επισκεφτεί την σελίδα. Από την σκοπιά της εξόρυξης γνώσης, χρησιμοποιούνται ως αντιπροσωπευτικά δείγματα των χρηστών, περιλαμβάνοντας πληροφορίες σχετικές με τις προτιμήσεις τους, τις σελίδες που επισκέπτονται, πόσο συχνά, και τι ενέργειες πραγματοποιούν.

Από την απαρχή της χρήσης τους στις αρχές του 2000, κατηγορήθηκαν από μεγάλες εφημερίδες όπως οι New York Times ως «εφαρμογές που οι εταιρείες εγκαθιστούν στον σκληρό δίσκο ώστε να υποκλέψουν προσωπικά δεδομένα από τον χρήστη», κάτι που δεν αληθεύει, μιας και ένα cookie περιλαμβάνει μόνο πληροφορίες που αποστέλλονται από την σελίδα κα αποθηκεύονται για μελλοντική χρήση. Κάθε cookie περιέχει επίσης ένα μοναδικό κωδικό αριθμό για λόγους ταυτοποίησης.

Η χρησιμότητα τους στην επιστήμη της εξόρυξης γνώσης είναι μεγάλη, εφόσον αποτελούν μία από τις κυριότερες πηγές συλλογής δεδομένων σε όλο το διαδίκτυο. Με τα cookies μπορούμε να γνωρίζουμε με μια ματιά πολλές πληροφορίες για την κίνηση των χρηστών σε μια ιστοσελίδα, τις περιοχές ενδιαφέροντος (για παράδειγμα το φόρουμ) αλλά και πιο σύνθετα συμπεράσματα όπως τον τόπο καταγωγής των χρηστών και την ακριβή θέση του δείκτη του ποντικιού τους.

Κύριο πλεονέκτημα της χρήσης τους, η εξατομίκευση, δηλαδή η αλλαγή των περιεχομένων της ιστοσελίδας ανάλογα με τις προτιμήσεις του χρήστη. Πρωτεργάτης αυτής της λειτουργίας, η Amazon, η οποία με την χρήση cookies μπορεί να καλωσορίσει τον χρήστη με το όνομά του, να του προβάλλει μόνο αντικείμενα σχετικά με τις προηγούμενες αναζητήσεις του αλλά και να «θυμάται» αντικείμενα που βρίσκονται στο καλάθι του ακόμα και αν κλείσει την σελίδα. Άλλες εταιρείες του χώρου δεν άργησαν να υιοθετήσουν την συνταγή, και σύντομα κάθε ιστοσελίδα μοίραζε cookies στους χρήστες.

Αυτή η τακτική έφερε πολλές συζητήσεις σχετικά με το τι δεδομένα καταγράφουν πραγματικά τα cookies και με τι σκοπό. Ορισμένοι χρήστες κατήγγειλαν την πολιτική αποδοχής cookies ως παραβίαση της ιδιωτικότητας, ενώ άλλοι δήλωσαν δυσαρεστημένοι με το γεγονός ότι συνέχισαν να δέχονται προσφορές σχετικές με προϊόντα που παρήγγειλαν πριν χρόνια. Επιπλέον, ερωτηματικά προκύπτουν από περιστατικά εμπορίας προσωπικών δεδομένων που καταγράφηκαν από cookies σε διαδικτυακά καζίνο και κόστισαν εκατοντάδες δολάρια στους κατόχους πιστωτικών καρτών.



Εικόνα 2.2.4: Ένα ειδικό πρόγραμμα ανιχνεύει μια λίστα με cookies εγκατεστημένα σε έναν υπολογιστή.

Από την πλευρά των επιχειρήσεων τώρα, τα cookies χρησιμοποιούνται και ως μέσο προώθησης νέων προϊόντων, δίνοντας για παράδειγμα εκπτώσεις σε πελάτες που επιστρέφουν συχνά (οι ιστοσελίδες που έχουν σχέση με τα ταξίδια χρησιμοποιούν αυτή την τεχνική για να δελεάσουν αναποφάσιστους πελάτες). Ακόμα, οι γίγαντες του διαδικτύου όπως η Microsoft, η Apple και η Google σχεδίασαν ειδικά για επαγγελματίες ειδικά cookies τα οποία βασίζονται πλέον στην μοναδική IP κάθε χρήστη, και επομένως είναι σχεδόν αδύνατον να απενεργοποιηθούν. Οι εταιρίες διαβεβαιώνουν ότι η συλλογή δεδομένων από αυτή την διαδικασία θα είναι πλήρως διαφανής και ότι ο χρήστης θα μπορεί να ελέγχει την ροή των δεδομένων του. Πολλοί χρήστες πάντως χρησιμοποιούν λογισμικό άρνησης cookies (Do Not Track) για να εμποδίσουν την εξαγωγή των στοιχείων τους.

Τα cookies ως τεχνολογία υπήρχαν από την αρχή της εξέλιξης του διαδικτύου και συνεχίζουν ακόμα και σήμερα να είναι από τους πιο αποδοτικούς τρόπους συλλογής δεδομένων για την εξαγωγή συμπερασμάτων. Η λογική της εξατομίκευσης της εμπειρίας του χρήστη σε συνδυασμό με την ευκολία χρήσης από λογισμικό εξόρυξης γνώσης κάνουν τα cookies ένα ισχυρό εργαλείο για κάθε επιχείρηση που θέλει να γνωρίζει περισσότερα για τους χρήστες της.

Επίλογος

Με βάση όλα τα προαναφερθέντα μέσα συλλογής δεδομένων, είναι σωστό να σημειωθεί ότι κανένα μέσο δεν είναι καλύτερο από τα άλλα και ότι σίγουρα δεν υπάρχει ένα για όλες τις δουλειές. Τόσο οι κλασικές όσο και οι σύγχρονες τεχνικές έχουν τα πλεονεκτήματα και τα μειονεκτήματά τους. Επίσης, επειδή χρησιμοποιούνται σε διαφορετικά μέσα, οι μεν κλασικές είναι πιο «χειροπιαστές» και απαιτούν πολλές φορές το πέρασμα δεδομένων με το χέρι, οι δε σύγχρονες κάνουν χρήση νέων τεχνολογιών και του διαδικτύου. Με την συνεργασία και των δύο, η επιστήμη της εξόρυξης γνώσης έχει πρόσβαση στην μεγαλύτερη δυνατή ποικιλία δεδομένων Επίλογος

προς συλλογή, ανάλυση, επεξεργασία και προβολή αποτελεσμάτων. Και το μέλλον θα δείξει τι καινούργιες τεχνολογίες θα εμφανιστούν και πως θα επεξεργάζονται τα δεδομένα. Στο επόμενο κεφάλαιο θα εξεταστούν οι τόποι που εφαρμόζεται η εξόρυξη γνώσης.

Κεφάλαιο 3: Τόποι Εφαρμογής Εξόρυξης Γνώσης

Εισαγωγή

Σε αυτό το κεφάλαιο θα μελετηθούν οι τόποι εφαρμογής των τεχνικών εξόρυξης γνώσης που αναλύθηκαν προηγουμένως. Διάφοροι τομείς της κοινωνίας έχουν διαφορετικές ανάγκες και ερμηνεύουν την ποιότητα των πληροφοριών που τους παρέχονται διαφορετικά. Θα μελετηθεί η εξέλιξη που έφερε η εξόρυξη γνώσης σε κάθε τομέα, από στρατιωτικά προγράμματα αντικατασκοπείας μέχρι και το πώς η δυνατότητα πρόβλεψης επηρεάζει την ποικιλία ραφιού στα σουπερ μάρκετ.

Ακόμη, θα αναδειχθούν οι ανάγκες που οδήγησαν κάθε τομέα στην αναζήτηση και χρήση τέτοιων αλγορίθμων καθώς και οι νέες απαιτήσεις που δημιουργήθηκαν μετά την έλευση των νέων τεχνολογιών και του διαδικτύου. Τέλος, θα υπάρξει μια σύντομη αναφορά σε εξειδικευμένες εφαρμογές που δημιουργήθηκαν ειδικά για συγκεκριμένους τόπους χρήσης.

3.1 Ιστοσελίδες

Ως εξόρυξη γνώσης σε ιστοσελίδες καλείται η χρήση των προαναφερθέντων τεχνικών στην μέση σελίδα, είτε αυτό είναι ένα blog είτε μια αθλητική κοινότητα είτε μια πανεπιστημιακή σχολή. Ακόμα και τα ιστολόγια με λιγότερη επισκεψιμότητα μπορεί να ωφεληθούν, καθώς μπορούν να αυξήσουν την δημοτικότητά τους με μερικές απλές αλλαγές.

Για παράδειγμα, ο ιδιοκτήτης ενός blog ειδήσεων παρατηρεί ότι η σελίδα έχει πολύ χαμηλή επισκεψιμότητα, αν και τα τελευταία νέα είναι πάντα αναρτημένα. Κάνοντας χρήση τεχνικών εξόρυξης γνώσεις (π.χ. στατιστική) βρίσκει πως οι περισσότεροι αναγνώστες προέρχονται από μια συγκεκριμένη χώρα, ή ότι ενδιαφέρονται για συγκεκριμένα θέματα περισσότερο από άλλα. Με αυτή την πληροφορία μπορεί να κάνει αλλαγές στον τρόπο οργάνωσης της σελίδας, αναπτύσσοντας περισσότερο προς την κατεύθυνση που θα φέρει περισσότερους αναγνώστες. Μια πανεπιστημιακή πύλη μπορεί να βρει με τον ίδιο τρόπο ποια βιβλία ζητούν περισσότερο οι φοιτητές, τα μαθήματα που πιθανόν χρειάζονται βοήθεια ή ακόμη και ποιοι μπορούν να βοηθήσουν με τα προβλήματα της εστίας.

Εκτός αυτού, οι τεχνικές εξόρυξης γνώσης βοηθούν την προβολή λιγότερο γνωστών σελίδων, οι οποίοι είναι μεν ποιοτικοί αλλά δεν διαθέτουν χρήματα για διαφήμιση. Εντοπίζοντας και προβάλλοντάς τους (συνήθως μέσω banner και affiliate rings) αυξάνεται και η ποιότητα του διαδικτύου ως σύνολο. Επίσης πρέπει να σημειωθεί ότι η όλη διαδικασία ανακάλυψης γνώσης είναι πλήρως αυτοματοποιημένη, και έτσι δεν υπάρχουν υποψίες για σαμποτάρισμα συγκεκριμένων σελίδων.

Συνήθως οι μεγάλες εταιρείες που εμπορεύονται αποθηκευτικό χώρο για ιστοσελίδες προσφέρουν έτοιμες επιλογές διαφήμισης για τα μέλη τους ανάλογα με το ποσό που διατίθενται να πληρώσουν. Συγκρίνοντας το περιεχόμενο της σελίδας με άλλες σχετικές μπορεί να δημιουργήσει affiliate rings και έτσι προσφέρουν καλύτερα οργανωμένο περιεχόμενο. Για παράδειγμα η εταιρεία web hosting GoDaddy προσφέρει συγκεκριμένα πακέτα τα οποία προβάλλουν την σελίδα του χρήστη σε περισσότερα δίκτυα που ανήκουν στην εταιρεία δωρεάν.

Τις περισσότερες φορές την εξόρυξη γνώσης την πραγματοποιούν οι web crawlers αυτοματοποιώντας την διαδικασία κατάταξης των ιστοσελίδων. Πολλές φορές όμως συναντούν εμπόδια, όπως ρήτρα απαγόρευσης ανίχνευσης, κακογραμμένο κώδικα, κατεστραμμένους τομείς ή νεκρούς συνδέσμους. Εκεί παρεμβαίνει είτε ο ίδιος ο χρήστης για να διορθώσει τα αποτελέσματα αυτά είτε η εταιρεία hosting με κάποια αυτοματοποιημένη λύση. Συνεχίζοντας το παράδειγμα της GoDaddy, συνεργάζεται με την DigitalCoconut, η οποία δημιουργεί λύσεις διαφήμισης με ειδικές εκπτώσεις για όσους χρησιμοποιούν ήδη την GoDaddy.

Από τα παραπάνω γίνεται κατανοητό ότι η επιστήμη της εξόρυξης γνώσης δεν προορίζεται μόνο για τους γίγαντες του κλάδου με εκατομμύρια επισκέπτες καθημερινά αλλά και για κάθε μικρή ιστοσελίδα που θέλει να αυξήσει την κίνησή της χρησιμοποιώντας τις προαναφερθείσες τεχνικές. Πολλές είναι οι περιπτώσεις που μικρές επιχειρήσεις εξελίχθηκαν σε διεθνείς κολοσσούς με την κατάλληλη προβολή. Χαρακτηριστικά παραδείγματα το Facebook και το Amazon, που αν και ξεκίνησαν ως πειραματικές σελίδες με ελάχιστο κοινό, είναι σήμερα από τις πιο προσοδοφόρες επιχειρήσεις. Το μόνο βέβαιο είναι πως αν υπάρχει ποιότητα, θα ανακαλυφθεί και θα προωθηθεί, και γρήγορα θα έρθει και η αναγνώριση.

3.2 Τράπεζες

Ίσως ο σημαντικότερος τομέας που έχει ανάγκη την εξόρυξη γνώσης είναι οι τράπεζες, τόσο για τις φυσικές όσο και για τις ηλεκτρονικές τους δοσοληψίες.

Οι τράπεζες θεωρούν ζωτικής σημασίας στόχο να γνωρίζουν τις προτιμήσεις και τις επιθυμίες των πελατών τους, προσαρμόζοντας ανάλογα τις υπηρεσίες τους. Τα σύγχρονα τραπεζικά συστήματα έχουν υιοθετήσει τις συναλλαγές μέσω διαδικτύου στην προσπάθεια βελτίωσης των συναλλαγών.

Ο λόγος που οι τράπεζες είναι δεκτικές προς τις τεχνικές εξόρυξης γνώσης είναι το ότι μέσα από την καθημερινή συναλλαγή με χιλιάδες πελάτες, υπάρχει μια στιβαρή βάση δεδομένων για ανάλυση και εξαγωγή συμπερασμάτων. Ακόμη, δεδομένα εισρέουν από ερωτηματολόγια και στατιστικές που πραγματοποιούν οι τράπεζες, ακόμα και από τα μέσα κοινωνικής δικτύωσης ή και απευθείας από τις τραπεζικές βάσεις δεδομένων. Αν για παράδειγμα μια μερίδα πελατών αντιμετωπίζει προβλήματα με τις διαδικτυακές συναλλαγές σε συγκεκριμένες ώρες, οι τεχνικές εξόρυξης γνώσης μπορούν να βοηθήσουν στην αναγνώριση του προβλήματος.

Ουσιαστικά οι λογαριασμοί των πελατών μελετούνται και αναλύονται, βοηθώντας την τράπεζα να δημιουργήσει νέα προϊόντα και υπηρεσίες. Αυτό όμως δεν είναι πάντα επιθυμητό, καθώς υπάρχει περίπτωση κάποιος πελάτης να έχει οικονομικά προβλήματα και να δεχτεί ένα εξειδικευμένο για αυτόν δάνειο, το οποίο στο μέλλον να αποδειχθεί μη συμφέρον λαμβάνοντας υπ' όψιν την νέα οικονομική κατάστασή του. Το πρόβλημα αυτό εμφανιζόταν κυρίως σε εποχές πριν την οικονομική κρίση, με τράπεζες να χρησιμοποιούν προ-εγκεκριμένα δάνεια και κάρτες οδηγώντας πολλούς καταναλωτές σε οικονομικό αδιέξοδο.

Παρ' όλα αυτά, υπάρχουν πολλά θετικά στοιχεία από την χρήση τεχνικών εξόρυξης γνώσης στον τραπεζικό χώρο. Η ασφάλεια των συναλλαγών έχει αυξηθεί κατακόρυφα. Έπειτα από απαιτήσεις των πελατών, νέα πρωτόκολλα ασφαλείας χρησιμοποιήθηκαν, διατέθηκαν κεφάλαια για την ανάπτυξη εφαρμογών ηλεκτρονικών τραπεζικών συναλλαγών και εκστρατείες ενημέρωσης ξεκίνησαν για να βοηθήσουν το κοινό να συνηθίσει στην χρήση του διαδικτύου για τις συναλλαγές του. Η αλήθεια είναι ότι οι Έλληνες δείχνουν να μην εμπιστεύονται την ηλεκτρονική τράπεζα, αν και τα ελληνικά συστήματα είναι ιδιαιτέρως ανταγωνιστικά.

Λόγω ύπαρξης ευαίσθητων προσωπικών δεδομένων σε ένα τραπεζικό λογαριασμό, η σχετική νομοθεσία προβλέπει την χρήση μόνο συγκεκριμένων για διαφημιστικούς σκοπούς, κυρίως οι κινήσεις που γίνονται σε έναν λογαριασμό. Αυτό και μόνο είναι αρκετό για τους ειδικούς, αφού γνωρίζοντας το μέσο ποσό ενός λογαριασμού μπορεί να εκτιμηθεί η αγοραστική δύναμη του χρήστη, οι καταναλωτικές του συνήθειες, ακόμη και το φύλο και η ηλικία του ή και πόσο συνεπής είναι στις οικονομικές του υποχρεώσεις.

Οι πληροφορίες αυτές είναι ζωτικές για την τράπεζα στην περίπτωση π.χ. που ζητηθεί ένα δάνειο μπορεί να γνωρίζει αν ο πελάτης θα αποπληρώσει τις δόσεις με συνέπεια. Αλλά και για να γνωρίζει τι ζητούν οι καταθέτες, τι πόρους πραγματικά έχει, τι κινήσεις κάνει ο ανταγωνισμός και έτσι να παραμένει κορυφαία στον τομέα της.

Από τα παραπάνω γίνεται κατανοητό ότι ο τραπεζικός χώρος έχει ανάγκη την εξόρυξη γνώσης. Είναι βέβαιο πως ακόμα και εν μέσω της οικονομικής κρίσης, καμιά τράπεζα δεν θα αρνούνταν να τις χρησιμοποιήσει για να αφουγκραστεί τις απαιτήσεις των πελατών της, και έτσι να εξελίσσεται συνεχώς και να παραμένει ανταγωνιστική στο σύγχρονο οικονομικό σκηνικό.

3.3 Παιχνίδια

Μία από τις λιγότερο προφανείς αλλά παρ' όλα αυτά σημαντική αγορά είναι ο χώρος της ανάπτυξης και πώλησης ηλεκτρονικών παιχνιδιών, είτε για υπολογιστές, είτε για κονσόλες. Οι εφαρμογές της εξόρυξης γνώσης δείχνουν εδώ τον δρόμο προς προσοδοφόρες επενδύσεις, μιας και η αγορά ηλεκτρονικών παιχνιδιών έχει την τάση να αλλάζει προτιμήσεις πολύ γρήγορα.

Η κυριότερη χρήση των τεχνικών εξόρυξης όπως και σε άλλους τομείς είναι η πρόβλεψη των αγοραστικών μοτίβων και διαθέσεων. Και με την αγορά να αλλάζει διαθέσεις συχνά, το να μπορεί μια εταιρεία να προβλέψει τις απαιτήσεις των πελατών πριν τις άλλες σημαίνει κέρδη εκατομμυρίων. Χαρακτηριστική η περίπτωση της Sega, η οποία στα τέλη του 1995 έδωσε στην αγορά την νέα της κονσόλα, το Sega Saturn χωρίς να έχει βολιδοσκοπήσει την ανταγωνιστική αγορά. Αποτέλεσμα ήταν η κονσόλα να αποτύχει παταγωδώς, κοστίζοντας 100 δολάρια περισσότερο από το PlayStation. Επίσης, έχοντας ως στόχο τις πωλήσεις των Χριστουγέννων η

Sega έδωσε το Saturn στην παραγωγή με ελάχιστη διαφήμιση και καθόλου αρχικούς τίτλους. Έπειτα από ζημιές εκατομμυρίων, η παραγωγή σταμάτησε. Αυτή ακριβώς η ανάγκη για γνώση και πρόβλεψη είναι κα η ουσία της εξόρυξης γνώσης. Αν και τα χρήματα που δαπανώνται για την διαφήμιση λογισμικού στην Ελλάδα είναι ελάχιστα σε σύγκριση με τα αντίστοιχα δεδομένα του εξωτερικού .

Οι τρόποι που η εξόρυξη γνώσης αντλεί δεδομένα από την αγορά ηλεκτρονικών παιχνιδιών δεν είναι οι παραδοσιακοί. Ελάχιστοι χρήστες θα αφιερώσουν χρόνο για να απαντήσουν σε ένα ερωτηματολόγιο ή μια έρευνα αγοράς. Έτσι δημιουργήθηκε η ανάγκη για την ανάπτυξη νέων τεχνικών. Ενέργειες του χρήστη όπως το αγοραστικό του προφίλ, δηλαδή το τι αγοράζει και κάθε πότε αγοράζει συλλέγονται και αντιπαραβάλλονται με άλλους χρήστες, δημιουργώντας έτσι μοτίβα συμπεριφοράς του καταναλωτικού κοινού ως σύνολο. Μια ακόμη δημοφιλής μέθοδος είναι η δημιουργία κοινοτήτων στις οποίες οι χρήστες μπορούν να συζητούν για τα παιχνίδια τους καθώς και να αποτυπώνουν την γνώμη τους.

Ακόμη, ο περιοδικός και ηλεκτρονικός τύπος στηρίζει την εξόρυξη γνώσης προσφέροντας επαγγελματικού τύπου δεδομένα μέσα από άρθρα, κριτικές και δημοσιεύσεις. Η δημοσιογραφία μπορεί να αναδείξει ένα νέο παιχνίδι και να βοηθήσει στην προβολή του σε όσο το δυνατόν μεγαλύτερη μερίδα του κοινού, με αναλύσεις και προβολές νέων τίτλων.

Ένα από τα προβλήματα που αντιμετωπίζει η διαδικασία άντλησης δεδομένων είναι το διαρκώς εναλλασσόμενο τοπίο το οποίο καλείται να εργαστεί. Προτιμήσεις που μέχρι πριν λίγες μέρες ήταν ορθές, αλλάζουν με το πέρασμα του χρόνου, την εξέλιξη της τεχνολογίας, ένα νέο μέσο (όπως η τεχνολογία αφής ή η χρήση νέων μέσων χειρισμού). Το κόστος όμως για την εξασφάλιση της συνεχούς ροής νέων στοιχείων για τις προτιμήσεις των καταναλωτών είναι ιδιαίτερα υψηλό. Σαν απάντηση, εταιρείες προσφέρουν προνόμια σε όσους συμμετέχουν ενεργά στις διαδικτυακές κοινότητες ή απαντούν σύντομα ερωτηματολόγια. Πολλοί χρήστες όμως απαντούν με ψευδή στοιχεία, ενδιαφερόμενοι μόνο για τα οφέλη που μπορούν να αποκομίσουν, χτίζοντας έτσι βάσεις δεδομένων με ψευδή στοιχεία. Τέλος, με το πάντα φλέγον ζήτημα της πειρατείας να μαστίζει τον κλάδο, χάνονται συνεχώς κεφάλαια που θα μπορούσαν να χρησιμοποιηθούν στην ανάπτυξη νέων, καλύτερων τίτλων και συρρικνώνοντας την αγορά.

Η αγορά των ηλεκτρονικών παιχνιδιών είναι το νέο μεγάλο στοίχημα της εξόρυξης γνώσης. Το τελικό αποτέλεσμα είναι αρκετά ικανοποιητικό σε σχέση με τα προηγούμενα χρόνια, όπου ο κλάδος αμφισβητούνταν από τους ειδικούς στον αν και κατά πόσο θα έχει μέλλον. Ήδη μεγάλες εταιρείες δημιουργούν δίκτυα για τους χρήστες τους (Steam, Origin, GOG, Xbox Live, PlayStation Network) όπου τα μέλη απολαμβάνουν προνόμια όπως ειδικά χαρακτηριστικά για τα παιχνίδια τους είτε ακόμα και πρόσβαση νωρίτερα από τους υπόλοιπους, αλλά και σημαντικές εκπτώσεις (Steam Sales, Origin, Early Access). Η συνεχής αλλαγή και προσαρμογή στις νέες απαιτήσεις είναι το κλειδί για την συνέχιση και την κερδοφορία των μεγάλων επιχειρήσεων του χώρου, αλλά και για την είσοδο των νέων δυναμικών εταιρειών με φρέσκες ιδέες, που διεκδικούν μερίδιο της αγοράς.

3.4 Εμπόριο

Ο κύριος λόγος ύπαρξης της ανάγκης για την εξόρυξη γνώσης είναι το εμπόριο. Η δημιουργία μοτίβων σχετικά με τις προτιμήσεις των καταναλωτών αλλά και η πρόβλεψη των μελλοντικών δεδομένων είναι μια άκρως απαραίτητη πληροφορία που κάθε επιχείρηση θα ήθελε να έχει.

Γιατί όμως να εφαρμόσει μια επιχείρηση τεχνικές εξόρυξης γνώσης; Μια εύκολη απάντηση είναι με σκοπό την μεγιστοποίηση του κέρδους. Υιοθετώντας για παράδειγμα ένα σύστημα εξαγωγής πληροφοριών όπως το Microsoft Analysis μπορεί να καταγράψει τις καταναλωτικές συνήθειες των πελατών και να εξάγει συμπεράσματα. Έπειτα χρησιμοποιώντας αυτή τη γνώση μπορεί να αναδιοργανώσει την δομή της, να εισάγει νέα προϊόντα και να γίνει πιο ανταγωνιστική.

Μια όχι και τόσο προφανής χρήση είναι το να μπορεί να γνωρίζει μια επιχείρηση πόσο αναγνωρίσιμη είναι και τι πιστεύει για εκείνη το καταναλωτικό κοινό (brand name). Υπάρχουν πολλές ιστοσελίδες που επιτρέπουν στους χρήστες να αφήνουν σχόλια σχετικά με τις εμπειρίες τους, και αντίστοιχα οι εταιρείες τα συμβουλεύονται για να εξελιχθούν.

Αλλά και η ίδια η επιχείρηση μπορεί να ωφεληθεί. Η έννοια της ανακάλυψης γνώσης δεν περιορίζεται μόνο στις στατιστικές. Χρησιμοποιώντας προγράμματα ανίχνευσης μπορεί για παράδειγμα να εντοπίσει υλικό και λογισμικό σε καλύτερες τιμές, ή ακόμα και νέες αγορές στις

οποίες θα μπορούσε να επεκταθεί. Κλασική η περίπτωση της Amazon, η οποία χρησιμοποιώντας τεχνικές μάρκετινγκ και εξόρυξης γνώσης (μέσα από στατιστικές και μοντέλα πρόβλεψης) έχει γίνει σήμερα το μεγαλύτερο ηλεκτρονικό κατάστημα παγκοσμίως.

Η άλλη όψη του νομίσματος είναι οι επιχειρήσεις που κάνουν εμπόριο προσωπικών δεδομένων έναντι αμοιβής χρησιμοποιώντας παράνομα μέσα εξόρυξης γνώσης όπως ιστοσελίδες με παραπλανητικό περιεχόμενο ή εγκαθιστώντας ανεπιθύμητο λογισμικό. Εταιρείες όπως η Conduit, MyWebSearch και Babylon ανταλλάσουν καθημερινά τα προσωπικά δεδομένα χιλιάδων χρηστών που κάνουν το λάθος να τις εμπιστευτούν. Πολλές φορές μάλιστα αρκεί ένα κλικ σε μια από τις διαφημίσεις που προβάλλουν για να εγκατασταθεί κρυφά. Κορυφαίες χώρες σε spamming και γενικότερα «ψάρεμα» δεδομένων είναι οι ΗΠΑ, η Κίνα και η Ρωσία, με τις κινέζικες εταιρείες hosting να κατέχουν τα πρωτεία σε κακόβουλες ιστοσελίδες λόγω της χαλαρής αντιμετώπισης τους από την κινεζική κυβέρνηση.

Η ποιότητα των δεδομένων που παράγουν τα λογισμικά εξόρυξης γνώσης είναι φυσικά τόσο καλή όσο τα στοιχεία που δίδονται για επεξεργασία. Είναι στατιστικά αποδεδειγμένο ότι οι χρήστες δέχονται να δώσουν προσωπικές πληροφορίες (όπως φύλο, ηλικία, προτιμήσεις) με αντάλλαγμα καλύτερες παρεχόμενες υπηρεσίες (προσωποποιημένα αποτελέσματα αναζήτησης, φιλτράρισμα προτάσεων που δεν ανταποκρίνονται στις προτιμήσεις τους). Συχνό φαινόμενο όμως είναι η ψευδής δήλωση στοιχείων ή ακόμη και η χρήση ειδικών εργαλείων παράκαμψης του βήματος εξολοκλήρου (όπως υπηρεσίες προσωρινού email, και ιστοσελίδες όπως το bugmenot, που προσφέρουν έτοιμους λογαριασμούς για δημοφιλείς σελίδες δωρεάν). Είναι προφανές ότι κανενός είδους γνώση δεν μπορεί να εξορυχτεί αν δεν μπορούμε να ταυτολογήσουμε τον χρήστη. Τέλος, αξίζει να σημειωθεί ότι υπάρχει μια διελκυστίνδα ανάμεσα στις επιχειρήσεις που θέλουν να γνωρίζουν όσο το δυνατόν περισσότερα και τους χρήστες που θέλουν να δώσουν όσο λιγότερα στοιχεία γίνεται. Για αυτό το λόγο θεσπίστηκε το 2003 ο νόμος προστασίας της ιδιωτικότητας στο διαδίκτυο (Online Privacy Protection Act – OPPA) ο οποίος υποχρεώνει κάθε επιχείρηση που συλλέγει προσωπικά δεδομένα να αναρτά σε εμφανές σημείο της ιστοσελίδας της την πολιτική προστασίας προσωπικών δεδομένων που ακολουθεί, καθώς και τον τρόπο που χειρίζεται τα στοιχεία που συλλέγει.

Οι επιχειρήσεις είναι η πλέον άμεση πλατφόρμα εφαρμογής της εξόρυξης γνώσης, με τους πελάτες να είναι η βάση και το λογισμικό να συνθέτει νέα μοτίβα τα οποία αξιοποιούνται για να βελτιωθεί η αγορά. Με την νόμιμη και ηθική χρήση των προσωπικών δεδομένων και την κατάλληλη επεξεργασία τους η αγορά οδηγείται σε νέες καινοτομίες που προκύπτουν από νέες ανάγκες.

3.5 Στρατός

Οι στρατιωτικές εφαρμογές της εξόρυξης γνώσης μπορεί να μην είναι άμεσα προφανείς, μιας και δεν εμπίπτουν στην λογική άντλησης πληροφοριών με σκοπό το κέρδος, αλλά περισσότερο με στόχο την ίδια την πληροφορία, η οποία μπορεί να κάνει την διαφορά. Με τα στρατιωτικά δεδομένα να αλλάζουν ταχύτατα η γνώση και η πρόβλεψη είναι άκρως σημαντικές έννοιες.

Οι απαρχές της σύγχρονης μιλιταριστικής εξόρυξης γνώσης έγιναν γνωστές στο ευρύ κοινό μετά από το σκάνδαλο των Αμερικανικών Μυστικών Υπηρεσιών το 2006 από τις οποίες διέρρευσαν βάσεις δεδομένων με προσωπικά στοιχεία χιλιάδων αμερικανών και μη πολιτών, τα οποία συλλέχθηκαν είτε από τα μέσα κοινωνικής δικτύωσης είτε από παρακολουθήσεις. Η κατακραυγή που ακολούθησε έκανε τους πολίτες καχύποπτους απέναντι στο κράτος το οποίο δεν μπορούσε ν εγγυηθεί την ιδιωτικότητα.

Τα δεδομένα που παράγονται παγκοσμίως είναι αστρονομικά σε μέγεθος. Καθημερινά η ανθρωπότητα παράγει 2,5 Exabyte (το ένα Exabyte είναι περίπου ένα εκατομμύριο terabytes). Αυτός ο όγκος δεδομένων περιέχει κλήσεις, μηνύματα, βίντεο, γραπτά κείμενα του μέσου χρήστη αλλά και αποτελέσματα πειραμάτων, απόρρητα σχέδια για επιχειρήσεις και συνομιλίες κατασκόπων. Ουσιαστικά οι αλγόριθμοι στρατιωτικού επιπέδου χαρτογραφούν το διαδίκτυο χρησιμοποιώντας «ετικέτες», δηλαδή δεδομένα για τα δεδομένα που περιγράφουν τι περιέχει μια ιστοσελίδα για να κρίνουν αν υπάρχει κάτι ενδιαφέρον. Ετικέτες όμως δεν υπάρχουν σε κάθε σελίδα (η NSA εκτιμά ότι μόλις το 3% του ορατού παγκόσμιου ιστού περιέχει αξιοποιήσιμες ετικέτες) και έτσι το έξυπνο λογισμικό τοποθετεί δικές του.

Οι ετικέτες αυτές ταυτοποιούν την σελίδα σε μία κλίμακα ενδιαφέροντος ανάλογα με τον αν περιέχει για παράδειγμα κείμενα που αφορούν την τρομοκρατία ή διακίνηση ναρκωτικών τότε τοποθετείται σε μία λίστα παρακολούθησης με σκοπό την εξαγωγή συμπερασμάτων για τον αν αποτελεί απειλή. Χαρακτηριστική η φράση αμερικανού αξιωματούχου για το πρόγραμμα ανάλυσης "Nexus 7" που χρησιμοποιήθηκε στον πόλεμο του Αφγανιστάν: «Το πρόγραμμα αναλύει τα πάντα, από εικόνες ραντάρ μέχρι διακυμάνσεις στις τιμές των φρούτων για να ανακαλύψουμε που κρύβονται τρομοκράτες»

Αυτά τα αποτελέσματα προωθούνται στις ειδικές ομάδες που ασχολούνται με την αντικατασκοπία και την παρακολούθηση και αναλύονται, το μεγαλύτερο μέρος από υπερυπολογιστές και αν εντοπιστεί μια απειλή από ανθρώπινο δυναμικό. Έχουν την δικαιοδοσία να ζητήσουν καταγραφές κλήσεων, κινήσεις λογαριασμών, μέχρι και ιατρικά δεδομένα, αν και οι σχετικές εταιρείες αρνούνται το ότι έχουν δώσει αυτές τις πληροφορίες.

Επίσης η αμερικανική (και όχι μόνο) κυβέρνηση έχει παραδεχθεί ότι στο παρελθόν χρησιμοποιούσε αναλυτές που παρακολουθούσαν τα μέσα κοινωνικής δικτύωσης και μάλιστα με υψηλές αμοιβές. Ιδιωτικές εταιρείες επέβλεπαν χιλιάδες «ευαίσθητα» άτομα, δηλαδή στόχους υψηλής προτεραιότητας και έστελναν τακτικά περιγραφές των κινήσεών τους. Αλλά και η κυβερνήσεις του Ισραήλ και της Κίνας έχουν στο δυναμικό τους επαγγελματίες των μέσων κοινωνικής δικτύωσης οι οποίοι προωθούν τα συμφέροντα της χώρας τους μέσω προπαγάνδας και ελέγχου.

Ένα σημαντικό ποσοστό των δεδομένων που καταγράφονται ως σημαντικά είναι λανθασμένοι συναγερμοί, κυρίως λόγω της αυτοματοποιημένης διαδικασίας η οποία αναζητά λέξεις-κλειδιά ανεξαρτήτως περιεχομένου. Χαρακτηριστική η περίπτωση όπου Γερμανοί αστυνόμοι των ειδικών δυνάμεων εισέβαλαν σε σπίτι φοιτητή λόγω του ότι το ιστορικό του φανέρωνε ότι έψαχνε επικίνδυνους όρους όπως «Τζιχάντ, Ισλάμ, Τρομοκρατία», όταν αποδείχθηκε ότι φοιτούσε στο τμήμα Μουσουλμανικών σπουδών του τοπικού πανεπιστημίου.

Οι στρατιωτικές εφαρμογές της εξόρυξης γνώσης ξεφεύγουν από την νόρμα της αναζήτησης πληροφοριών με σκοπό το κέρδος. Σκοπός είναι η διαφύλαξη της ειρήνης και της προστασίας του κοινού από εξωτερικές και εσωτερικές απειλές. Στην προσπάθεια να επιτύχει αυτός ο σκοπός πολλές φορές η ιδιωτικότητα παραβιάζεται. Κάποιοι πιστεύουν ότι είναι αναγκαίο κακό

ενώ άλλοι παλεύουν για να παραμείνει το απαράβατο της ιδιωτικής ζωής. Ένα είναι σίγουρο όμως: Η εξόρυξη γνώσης και τα εργαλεία της θα συνεχίσουν να αποτελούν κομμάτι του στρατού για πολύ καιρό ακόμα.

Επίλογος

Σε αυτό το κεφάλαιο μελετήθηκαν ορισμένοι από τους σημαντικότερους τόπου όπου εφαρμόζεται η εξόρυξη γνώσης. Ακόμα αναλύθηκαν τα ειδικά προβλήματα που αντιμετωπίζει κάθε κλάδος, από την τοποθέτηση των προϊόντων στα ράφια ενός σούπερ μάρκετ ανάλογα με τις προτιμήσεις των πελατών μέχρι τους ισχυρούς αλγόριθμους που διασφαλίζουν τις τραπεζικές συναλλαγές αλλά και τις ανάγκες του στρατού για πληροφόρηση.

Κάθε χώρος εφαρμογής αξιοποιεί την γνώση που αποκομίζει με διαφορετικό τρόπο, και όχι πάντα με σκοπό το κέρδος. Η κοινή τους όμως περιοχή είναι η ανάγκη για γνώση, για εξαγωγή μοτίβων και συμπεριφορών και για πρόβλεψη των μελλοντικών δεδομένων τους. Το διαδίκτυο περιέχει πληθώρα πληροφοριών αλλά και πολύ θόρυβο και εναπόκεινται στην επιστήμη της εξόρυξης γνώσης να τα διαχωρίσει. Στο επόμενο κεφάλαιο θα αναλυθούν ορισμένες από τις κυριότερες εφαρμογές της εξόρυξης γνώσης που συνδέονται με την αγορά του διαδικτύου.

Κεφάλαιο 4: Εφαρμογές Εξόρυξης Γνώσης

Εισαγωγή

Στο παρόν κεφάλαιο θα μελετηθούν ορισμένες από τις κυριότερες εφαρμογές της επιστήμης της εξόρυξης γνώσης. Από την δημιουργία στοχευόμενων διαφημίσεων έως και την στατιστική ανάλυση μεγάλων βάσεων δεδομένων, ο κλάδος έχει να προσφέρει μια λύση σε κάθε ανάγκη για γνώση και πρόβλεψη. Επίσης, θα γίνει αναφορά σε ορισμένες γνωστές τεχνικές μάρκετινγκ που σχετίζονται άμεσα με την εξόρυξη γνώσης όπως τα αγοραστικά προφίλ και η δημιουργία εξατομικευμένων προτάσεων αγορών.

Ουσιαστικά, οι εφαρμογές της εξόρυξης γνώσης είναι απεριόριστες. Χρησιμοποιώντας τις κατάλληλες τεχνικές και ένα σύνολο δεδομένων προς επεξεργασία ο καθένας μπορεί να συνθέσει νέα συμπεράσματα και να προβλέψει μελλοντικά. Εδώ όμως θα αναλυθούν ειδικευμένες εφαρμογές οι οποίες χρησιμοποιούνται από κολοσσούς του διαδικτύου και έχουν εξελιχθεί σε πυλώνες της επιχειρηματικότητας και όχι μόνο. Ας τις εξετάσουμε.

4.1 Αγοραστικά Προφίλ

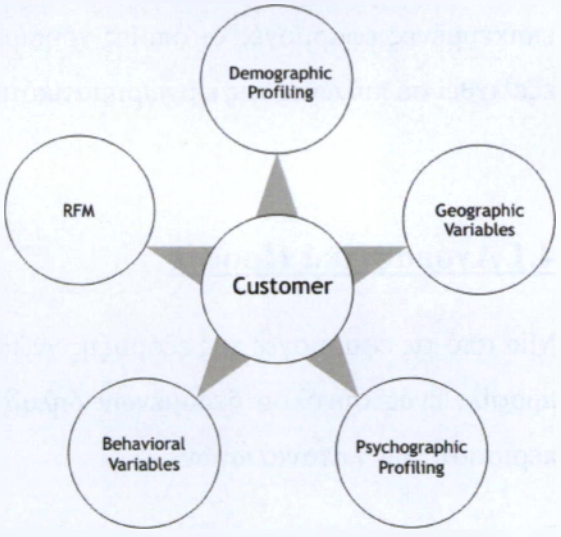
Μία από τις εφαρμογές της εξόρυξης γνώσης στο διαδίκτυο είναι και η δημιουργία αγοραστικών προφίλ, ενός συνόλου δεδομένων δηλαδή που αφορούν στοιχεία για τις προτιμήσεις ενός ή περισσότερων καταναλωτών.

Τι σημαίνει η δημιουργία καταναλωτικού προφίλ για τον ίδιο τον χρήστη; Κυρίως μια διαφοροποιημένη εμπειρία στις αγοραστικές του συνήθειες, με εξατομικευμένες προτάσεις, εκπτώσεις και προσφορές σε είδη που τον ενδιαφέρουν αλλά και ευκολίες όπως το να θυμάται η σελίδα το καλάθι των αγορών του, να αναγνωρίζει το όνομά του και να τον ειδοποιεί για νέα αντικείμενα ακόμα και αν βρίσκεται εκτός της σελίδας. Επιπλέον, πολλοί καταναλωτές έχουν δηλώσει ότι είναι διατεθειμένοι να δώσουν προσωπικά τους στοιχεία με αντάλλαγμα καλύτερη εξυπηρέτηση.

Από την σκοπιά των επιχειρήσεων, ένα αγοραστικό προφίλ φτιάχνεται δυσκολότερα, αλλά περιέχει περισσότερα στοιχεία. Η εταιρία ξεκινά αρχικά με ένα σύνολο «ιδανικών αγοραστών»,

δηλαδή εκείνων που έχουν ανάγκη το προϊόν της εταιρείας, και είναι δεκτικοί σε πωλήσεις. Έπειτα, προσθέτονται μοναδικά στοιχεία που κάνουν κάθε ομάδα υποψηφίων αγοραστών μοναδική όπως η γλώσσα που μιλάνε, τι τους εμποδίζει να αγοράσουν το προϊόν και η ιδανική μέθοδος προσέγγισης για επιτυχημένες πωλήσεις. Το επόμενο βήμα για την δημιουργία του προφίλ είναι να ανακαλύψει η επιχείρηση ποια μέσα χρησιμοποιούν οι πελάτες για να μάθουν για την επιχείρηση. Τα κοινωνικά δίκτυα; Την επίσημη ιστοσελίδα; Ποιες λέξεις κλειδιά οδηγούν στην σελίδα και πόσο εύκολο είναι να εντοπιστεί από τις μηχανές αναζήτησης; Αυτές οι απαντήσεις θα δημιουργήσουν ένα ολοκληρωμένο αγοραστικό προφίλ, το οποίο η επιχείρηση μπορεί να μετατρέψει σε επιτυχημένες πωλήσεις κάνοντας τις απαραίτητες αλλαγές.

Ένα παράδειγμα αυτής της λογικής είναι μια επιχείρηση που εμπορεύεται κάρτες ήχου. Έπειτα από έρευνα διαπιστώνεται ότι το αγοραστικό κοινό δεν ανταποκρίνεται στις διαφημιστικές καμπάνιες και οι πωλήσεις είναι χαμηλές. Αναζητώντας το προφίλ του ιδανικού αγοραστή, η εταιρία ανακαλύπτει ότι το 80% των πελατών της βρίσκονται στο Ηνωμένο Βασίλειο, και οι περισσότεροι είναι μουσικοί παραγωγοί. Ο λόγος που δεν προτιμούν την εταιρεία φαίνεται να είναι η έλλειψη πληροφόρησης, καθώς βρίσκεται αρκετά χαμηλά στις μηχανές αναζήτησης. Από τα παραπάνω, το αγοραστικό προφίλ του ιδανικού πελάτη είναι ο Βρετανός μουσικός παραγωγός και η καλύτερη μέθοδος προσέγγισης είναι η στοχευόμενη διαφήμιση. Με την γνώση αυτή είναι σίγουρο ότι οι πωλήσεις θα αυξηθούν.



Εικόνα 4.1: Απαραίτητα στοιχεία για την δημιουργία αγοραστικών προφίλ.

Οι πολέμιοι των αγοραστικών προφίλ θεωρούν ότι η θίγεται η ιδιωτικότητα του ατόμου αναγκάζοντας τον καταναλωτή να δώσει προσωπικά στοιχεία για να έχει πρόσβαση σε προνόμια. Επιπλέον, συχνά αναφέρεται η τάση των εταιρειών να αποκτούν στοιχεία χωρίς την άδεια των πελατών με την χρήση cookies αλλά και αθέμιτων μέσων όπως το ηλεκτρονικό ψάρεμα (phishing). Επίσης, τα αγοραστικά προφίλ υλοποιούνται συνήθως από έναν μέσο όρο των καταναλωτών, πράγμα που σημαίνει ότι κάποιος πελάτης που δεν εντάσσεται στο «στοχευόμενο» κοινό ίσως αντιμετωπίσει δυσκολίες στις αγορές του (για παράδειγμα μια

ιστοσελίδα με κύριους πελάτες Γερμανούς αποφασίζει να αλλάξει την γλώσσα της ιστοσελίδας τους στα γερμανικά, αποκλείοντας έτσι τους αγγλόφωνους πελάτες). Τέλος, ένα αγοραστικό προφίλ δεν έχει καμία αξία αν δεν υπάρχει ανάλογη αρχιτεκτονική που να κάνει χρήση της γνώσης που αποκομίστηκε. Για παράδειγμα, το να γνωρίζει μια εταιρεία ότι ο μέσος καταναλωτής της επιθυμεί καλύτερες υπηρεσίες στην χώρα του δεν σημαίνει τίποτε αν δεν υπάρχουν αντίστοιχα υποκαταστήματα.

Τα αγοραστικά προφίλ είναι μια μη συμβατική μορφή εξόρυξης γνώσης, προσπαθώντας να συνθέσει τον «ιδανικό πελάτη» με βάση τις προτιμήσεις του συνόλου των ήδη υπαρχόντων. Κάνοντας χρήση αυτής της γνώσεις οι επιχειρήσεις εξελίσσονται συνεχώς, αλλάζοντας τα προϊόντα και τις υπηρεσίες τους σύμφωνα με τις ανάγκες και τις προτιμήσεις των καταναλωτών τους. Η προστασία της ιδιωτικότητας είναι πάντα ένα σημαντικό θέμα, και κάθε πελάτης θέλει εγγυήσεις, όμως πάντα θα υπάρχει το δίλημμα ανάμεσα στην ιδιωτικότητα χωρίς προσφορές και τις προσφορές με αντάλλαγμα προσωπικά δεδομένα.

4.2 Στατιστικές Αναλύσεις

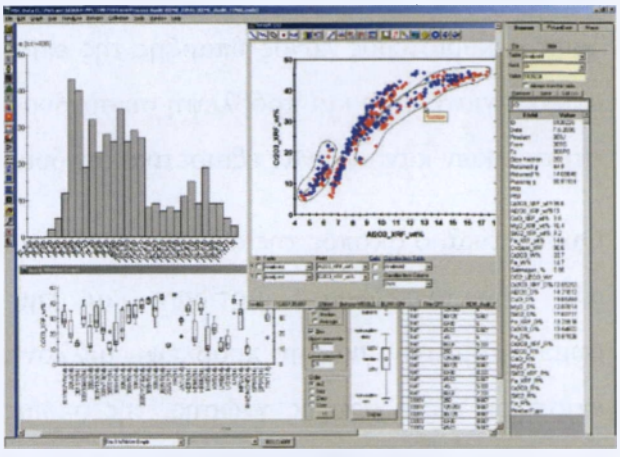
Ίσως ο κυριότερος λόγος ύπαρξης της επιστήμης της εξόρυξης γνώσης είναι η στατιστική. Η ανάγκη για γνώση και πρόβλεψη σε συνδυασμό με την ευκολία παρουσίασης και ανάλυσης των στατιστικών κάνουν έναν εξαιρετικό συνδυασμό.

Ουσιαστικά, ο σκοπός της στατιστικής είναι η μελέτη των δεδομένων. Οι χρήσεις της ποικίλουν από την αναγνώριση μοτίβων και την απεικόνιση γραφημάτων μέχρι την εξαγωγή συμπερασμάτων για την πρόβλεψη μελλοντικών εισόδων, κάτι με το οποίο ασχολείται και η επιστήμη της εξόρυξης γνώσης, της οποίας ο σκοπός είναι η επεξεργασία των εισόδων με κατάλληλο τρόπο ώστε να προκύψουν νέα και πρωτότυπα συμπεράσματα. Η ποιοτική διαφορά της εξόρυξης γνώσης από την στατιστική όμως, είναι η ικανότητα της πρώτης να φιλτράρει τον θόρυβο, δηλαδή άσχετα δεδομένα, αλλά και να συνθέσει αποδεκτά αποτελέσματα για να αντικαταστήσει εγγραφές που λείπουν (σύνηθες φαινόμενο σε βάσεις δεδομένων).

Οι πέντε βασικές τεχνικές της εξόρυξης γνώσης, δηλαδή η διαχώριση, ομαδοποίηση, πρόβλεψη, ανάλυση και παρουσίαση εφαρμόζονται και στην στατιστική, με διαφορετική φιλοσοφία. Στην εξόρυξη γνώσεις, οι τρεις πρώτες βρίσκουν έδαφος κυρίως στην μηχανική μάθηση υπό επίβλεψη, δηλαδή το να εξελίσσεται ένα σύστημα με την βοήθεια ενός σετ δοκιμαστικών δεδομένων με την ανθρώπινη υποβοήθηση στον να κατανοήσει ποια είναι σωστά. Η μη επιβλεπόμενη μηχανική μάθηση έκανε χρήση κυρίως των τριών τελευταίων, συνθέτοντας νέα δεδομένα με βάση τα παλαιότερα και προβλέποντας τα μελλοντικά.

Η επιστήμη της στατιστικής χρησιμοποιεί τις ίδιες τεχνικές, με διαφορετικό τρόπο. Κατ' αρχάς, η στατιστική βασίζεται σε συμπαγή δεδομένα (χωρίς θόρυβο, ελλείψεις ή λανθασμένα δεδομένα) για να αναπαραστήσει και να προβλέψει μελλοντικά αποτελέσματα. Η ικανότητα επεξεργασίας των εγγραφών πριν την παρουσίαση είναι αυτό που διαχωρίζει τις δύο επιστήμες. Ουσιαστικά, η στατιστική και η εξόρυξη γνώσης είναι αλληλένδετες, με κοινό στόχο και διαφορετικά εργαλεία.

Για παράδειγμα, μια στατιστική μπορεί να απεικονίσει τις τάσεις της αγοράς σε σχέση με ένα νεοεμφανιζόμενο προϊόν και να καταγράψει το πόσο επιτυχημένο είναι. Προσθέτοντας τώρα τις τεχνικές εξόρυξης γνώσης μπορούμε να μάθουμε τις ηλικίες που απευθύνεται, τις χώρες που είναι δημοφιλέστερο, τον μέσο όρο ζωής του προϊόντος, και τι μπορεί να αλλάξει ώστε να γίνει πιο επιτυχημένο. Ενώνοντας λοιπόν τις δύο επιστήμες έχουμε την μεγιστοποίηση της κατανόησης των αποτελεσμάτων και την ικανότητα να προβλέψουμε τα μελλοντικά.



Εικόνα 4.2 Ένα πρόγραμμα επεξεργασίας στατιστικών αναλύει ένα γράφημα

Από την στατιστική όμως μπορεί να επωφεληθεί και ο καταναλωτής, διευρύνοντας τις επιλογές του και λαμβάνοντας στοιχεία για να συγκρίνει τιμές και χαρακτηριστικά των προϊόντων που τον ενδιαφέρουν. Επιπλέον, πολλοί καταναλωτές διενεργούν δικές τους στατιστικές, τις οποίες μεγάλες εταιρείες συμβουλεύονται καθώς τα δεδομένα που προκύπτουν είναι πάντα σχετικά με τα ενδιαφέροντά τους. Χαρακτηριστική η περίπτωση της Lego, η οποία επέτρεψε στους πελάτες της να διενεργήσουν δικές τους έρευνες και να αποφασίσουν τι νέο θα κατασκεύαζε η εταιρεία.

Η έρευνα ολοκληρώθηκε με επιτυχία και από τα αποτελέσματά της δημιουργήθηκε η σειρά EV3, την νικητήρια πρόταση μάλιστα την κατέθεσε Έλληνας. Η Lego αποκόμισε δωρεάν διαφήμιση αφενός, δωρεάν έρευνα αγοράς που φανέρωσε τις προτιμήσεις των καταναλωτών της, αλλά και βελτίωσε την εικόνα της στα μάτια των καταναλωτών, δείχνοντας ότι ακούει τις απόψεις τους. Η στατιστική αυτή έρευνα πραγματοποιείται πλέον κάθε χρόνο.

Η εξόρυξη γνώσης και η στατιστική είναι δύο όψεις του ίδιου νομίσματος, κάνοντας την κατανόηση, ανάλυση και παρουσίαση δεδομένων ευκολότερη για τον άνθρωπο. Σύνθετα αποτελέσματα μπορούν να ανακαλυφθούν με τις τεχνικές της εξόρυξης γνώσης και να γίνουν πιο κατανοητά και στον επιχειρηματία αλλά και στον μέσο χρήστη. Η στατιστική είναι μία από τις σημαντικότερες εφαρμογές της εξόρυξης γνώσης και ένας από τους πυλώνες του σύγχρονου μάρκετινγκ.

4.3 Διαφήμιση

Ένας από τους κυριότερους λόγους χρήσης τεχνικών εξόρυξης γνώσης είναι και η στοχευόμενη διαφήμιση. Ως στοχευόμενη διαφήμιση καλείται η οπτικοακουστική παρουσίαση προσφορών στους χρήστες με συγκεκριμένες προτιμήσεις.

Η αγορά του διαδικτύου βρίθει διαφημίσεων ποικίλης ποιότητας. Σε ένα χώρο όπου ο καθένας μπορεί να εισέλθει και να διεκδικήσει μερίδιο της αγοράς ο ανταγωνισμός είναι λυσσαλέος. Μία διαφήμιση στοχευόμενη στον σωστό αποδέκτη μπορεί να κάνει την διαφορά ανάμεσα στο σωστό μάρκετινγκ και την συνεχή και άσχετη ανεπιθύμητη αλληλογραφία. Εδώ εισέρχονται οι τεχνικές εξόρυξης γνώσης, με τις οποίες μπορεί να γίνει αντιληπτό το μέσο πλήθος επισκεπτών κάθε σελίδας, καθώς και τα ενδιαφέροντά τους. Έτσι, προβάλλονται διαφημίσεις και προσφορές που έχουν μεγαλύτερη πιθανότητα να προσελκύσουν τους χρήστες. Για παράδειγμα, banners με διαφημίσεις καλλυντικών θα έχουν μικρή απήχηση σε μια ιστοσελίδα που αφορά τις πολεμικές τέχνες. Καλύτερη επιλογή θα ήταν γάντια πυγμαχίας, σάκοι προπόνησης ή ενεργειακά ποτά.

Υπάρχουν πολλοί τρόποι για να προσδιοριστεί το κοινό μιας σελίδας. Συνήθως η διαδικασία αυτοματοποιείται χρησιμοποιώντας web crawlers, οι οποίοι ανιχνεύουν μια ιστοσελίδα για

«ετικέτες», κομμάτια κώδικα δηλαδή που αναφέρουν λέξεις κλειδιά σχετικά με τα περιεχόμενα. Έπειτα τα δεδομένα αυτά ομαδοποιούνται σε μεγάλες κατηγορίες όπως άνδρες, αθλητικά, μοντελισμός, παπούτσια και συνδυάζονται με έτοιμες διαφημίσεις με παρόμοιες «ετικέτες» για τη αποδοτικότερη προβολή τους.

Αν και ο μηχανισμός λειτουργίας είναι απλός, εντούτοις υπάρχουν και ορισμένα προβλήματα όπως διαφημιστές να χρησιμοποιούν «ετικέτες» άσχετες με το θέμα που διαφημίζουν, εκμεταλλευόμενοι την αυτοματοποιημένη διαδικασία. Τέτοιες διαφημίσεις όμως γρήγορα αποσύρονται καθώς καταγγέλλονται από τους ιδιοκτήτες των σελίδων. Επιπλέον, συχνό φαινόμενο είναι banners τα οποία είναι ιδιαίτερα ενοχλητικά παίζοντας δυνατή μουσική ή βίντεο, εικόνες πορνογραφικού περιεχομένου σε άσχετες σελίδες είτε τέλος με την τοποθέτησή τους σε ζωτικά σημεία της ιστοσελίδας όπως το πλήκτρο λήψης, αναγκάζοντας τον χρήστη να τις επισκεφθεί.

Η λογική μιας διαφήμισης είναι φυσικά να προσελκύσει τον αγοραστή. Οι σύγχρονες τεχνικές εξόρυξης γνώσης κάνουν την διαφήμιση να εμφανίζεται ταυτόχρονα σε χιλιάδες διαφορετικές σελίδες, οι οποίες τις περισσότερες φορές πληρώνονται για την προβολή τους. Υπάρχουν δύο διαφορετικές οδοί πληρωμής: η πληρωμή κάθε φορά που ο χρήστης κάνει κλικ στην διαφήμιση και η πληρωμή για την απλή προβολή της. Αλλά και εκτός διαδικτύου, κορυφαία τηλεοπτικά δίκτυα όπως το βρετανικό Sky TV χρησιμοποιούν τις έξυπνες τηλεοράσεις για να αντλήσουν δεδομένα προτίμησης των τηλεθεατών όπως σειρές που βλέπουν, εμφανίζοντας κατάλληλα τροποποιημένες διαφημίσεις, εξειδικευμένες για κάθε μοναδικό τηλεθεατή. Ωστόσο, πολλά δίκτυα αρνούνται να κάνουν χρήση αυτής της τεχνολογίας προβάλλοντας ως αντεπιχείρημα την παραβίαση της ιδιωτικότητας των τηλεθεατών.

Εικόνα 4.3: Προτάσεις αγοράς με βάση προηγούμενα στοιχεία από την Amazon.

Μια άλλη ενδιαφέρουσα πτυχή της στοχευόμενης διαφήμισης είναι η χρήση έξυπνων αλγόριθμων οι οποίοι δεν περιορίζονται στην απλή ταυτοποίηση «ετικετών», αλλά συλλέγουν

δεδομένα σε μεγάλες ποσότητες (και με την εύνοια των γιγάντων του διαδικτύου όπως η Google που τα προσφέρει δωρεάν), δημιουργώντας λογικές σχέσεις ανάμεσα σε έννοιες, κάνοντας δυνατή την αποστολή ειδικευμένων διαφημίσεων σε χρήστες χωρίς οι ίδιοι να έχουν επισκεφτεί μια σχετική σελίδα. Για παράδειγμα, αν χιλιάδες χρήστες αρχίζουν να αναζητούν συγκεκριμένες μάρκες κινητών τηλεφώνων και στην συνέχεια φορτιστές, νέοι πελάτες που αναζητούν τις ίδιες συσκευές θα βλέπουν διαφημίσεις για φορτιστές χωρίς να τους αναζητούν, με βάση προηγούμενες συμπεριφορές άλλων χρηστών. Η μηχανική αυτή μάθηση είναι η κορυφαία τεχνολογία στον χώρο του μάρκετινγκ.

Η διαφήμιση είναι ένας από τους πυλώνες της εξόρυξης γνώσης και μία από τις πιο προφανείς χρήσης της. Η δυνατότητα μεγιστοποίησης του κέρδους με την εφαρμογή μερικών απλών βημάτων κάνει το μήνυμα της διαφήμισης να φτάνει περισσότερους καταναλωτές, σε περισσότερες σελίδες, με λιγότερη προσπάθεια. Η εξ' ολοκλήρου αυτοματοποίηση της διαδικασίας έχει θετικές και αρνητικές πλευρές, αλλά σαν σύνολο λειτουργεί και θα συνεχίσει έτσι για πολύ καιρό ακόμα.

4.4 Ασφάλεια

Ο τομέας της ασφάλειας, είτε πρόκειται για την ασφάλεια των ηλεκτρονικών συναλλαγών είτε για τραπεζικά συστήματα είτε τέλος για την ασφάλεια του ιδιωτικού απόρρητου είναι ένας από τους στόχους της επιστήμης της εξόρυξης γνώσης.

Η έννοια της ασφάλειας έχει διαφορετική σημασία γα κάθε κλάδο. Για παράδειγμα, ασφάλεια για μια ηλεκτρονική συναλλαγή θα ήταν η εγγύηση ότι κανείς δεν θα μπορεί να υποκλέψει τα στοιχεία του πελάτη. Ασφάλεια για μια εταιρεία παροχής ηλεκτρονικών υπηρεσιών θα ήταν τα δεδομένα της να παραμείνουν μακριά από εισβολείς. Τέλος, για τον μέσο χρήστη σημαίνει ότι η ιδιωτικότητα του δεν θα παραβιάζεται και το να μπορεί να μην αισθάνεται απειλούμενος ενώ βρίσκεται στο διαδίκτυο.

Η εξόρυξη γνώσης λειτουργεί και από τις δύο όψεις του νομίσματος, δηλαδή και από την πλευρά που θέλει να γνωρίζει περισσότερα, αλλά και από εκείνη που προσπαθεί να προστατέψει

μαθαίνοντας και προβλέποντας νέους κινδύνους που εμφανίζονται. Ήδη υπάρχουν στοιχεία που δείχνουν ότι κυβερνήσεις μεγάλων κρατών έχουν συνάψει συμφωνίες με γίγαντες του διαδικτύου όπως η Microsoft και η Google για την απόδοση προσωπικών στοιχείων σε περιπτώσεις που ζητηθεί για λόγους εθνικής ασφάλειας, Από την άλλη μεριά οι χρήστες θέλοντας πάντα να παραμένουν όσο το δυνατόν ανώνυμοι έχουν καταφέρει να βρουν «τρύπες» στο σύστημα αλλά και στην νομοθεσία έτσι ώστε να αποφύγουν την καταγραφή. Ήδη από το 2004 υπάρχει νόμος που προστατεύει την ανωνυμία του διαδικτύου και τα προσωπικά δεδομένα.

Ασφάλεια όμως σημαίνει και γνώση. Η νομοθεσία αναφέρει ότι απαγορεύεται να μάθει κάποιος περισσότερα για κάποιον μόνο με την εφαρμογή τεχνικών εξόρυξης γνώσης, παρά μόνο να συνθέσει λογικά αποτελέσματα με βάση παρατηρήσεις. Αν και κάθε έξυπνος αλγόριθμος μπορεί να προβάλλει μοτίβα σχετικά με τις προτιμήσεις των καταναλωτών, είναι δύσκολο να αποδώσει ακριβή προσωπικά στοιχεία για κάθε ξεχωριστό άτομο.

Για παράδειγμα, η εξόρυξη γνώσης εγγυάται έναν βαθμό ασφάλειας σε μια βάση δεδομένων ψάχνοντας για εγγραφές που είναι εκτός της νόρμας και σε συνδυασμό με ένα λογισμικό προστασίας μπορεί να απομονώσει απειλές μέσω της ικανότητας πρόβλεψης. Επιπλέον, προγράμματα ανίχνευσης διατρέχουν τον ιστό συλλέγοντας πληροφορίες για ιστότοπους που «ψαρεύουν» προσωπικά δεδομένα και ανανεώνουν συνεχώς τις λίστες τους, προσφέροντας ένα ασφαλέστερο διαδίκτυο για όλους.

Αξίζει να σημειωθεί εδώ ότι η χρήση προγραμμάτων εξόρυξης γνώσης με σκοπό την διαδικτυακή έχει ορισμένα σημαντικά μειονεκτήματα, συγκεκριμένα τα λανθασμένα θετικά και τα λανθασμένα αρνητικά αποτελέσματα. Λανθασμένο θετικό έχουμε όταν μια σελίδα ή μια εγγραφή επισημαίνεται ως επικίνδυνη ενώ δεν είναι. Αυτό οφείλεται πολλές φορές σε λάθος ρυθμίσεις του αλγόριθμου ανίχνευσης, είτε στην ίδια την εγγραφή η οποία μπορεί να μην είναι επικίνδυνη, απλώς λανθασμένη. Τέλος, ψευδές αρνητικό έχουμε όταν εγγραφές ή σελίδες που θα ήταν επικίνδυνες παραβλέπονται είτε γιατί αποκρύπτουν την κακόβουλη ιδιότητά τους από το πρόγραμμα ανίχνευσης είτε πρόκειται για μια σελίδα που αντικαταστάθηκε από κακόβουλο λογισμικό αφού της δόθηκε πιστοποιητικό ασφαλείας (σύνηθες φαινόμενο σε banners που διαφημίζουν ηλεκτρονικές προσφορές).

Αυτά τα προβλήματα αντιμετωπίζονται με βελτίωση των πρωτοκόλλων ασφαλείας των αλγορίθμων εξόρυξης, μια διαδικασία που μπορεί να είναι είτε επανάληψη του δοκιμαστικού σετ δεδομένων είτε ακόμα προσθήκη δικλείδων ασφαλείας με υψηλότερες προδιαγραφές. Για παράδειγμα, ένα πρόγραμμα ανίχνευσης «διαβάζει» την ώρα εισόδου και την πύλη που χρησιμοποίησε ο χρήστης για να εισέλθει στο σύστημα. Αν μόνο οι πύλες 86 και 439 έχουν πιστοποιητικό ασφαλείας, κάθε άλλη περίπτωση θα θεωρείται από το σύστημα ως παραβίαση, κάτι που μπορεί να είναι αναληθές. Βελτίωση αυτού του σεναρίου θα ήταν μια καλύτερη ενημέρωση του συστήματος σχετικά με όλες τις ανοικτές πύλες καθώς και την βαρύτητα της καθεμίας, έτσι ώστε να μπορεί να αποφασίσει αν μια νέα είσοδος είναι κακόβουλη ή όχι.

Η ασφάλεια είναι μία από τις βασικές ανάγκες του σύγχρονου κόσμου του διαδικτύου. Με την αγορά να ανταγωνίζεται συνεχώς με θεμιτά και αθέμιτα μέσα, με επιτιθέμενους και αμυνόμενους, και με τον απλό χρήστη να θέλει εγγυήσεις για την ιδιωτικότητα του, η εξόρυξη γνώσης μπορεί να θέσει τις βάσεις για την βελτίωση των ήδη υπαρχόντων κανόνων ασφαλείας και την βελτίωση του πεδίου μάχης που λέγεται διαδίκτυο. Γιατί στην τελική, ασφάλεια σημαίνει γνώση, και η γνώση φέρνει την επιτυχία.

4.5 Εμπορικές Εφαρμογές

Σε αυτή την ενότητα θα αναλυθούν ορισμένες από τις γνωστότερες εμπορικές εφαρμογές που σχετίζονται με την εξόρυξη γνώσης. Καθένας από τους τρεις τομείς της (στατιστική, μηχανική μάθηση και τεχνητή νοημοσύνη) έχουν στο ενεργητικό τους προγράμματα αξίας εκατοντάδων ευρώ αλλά και δωρεάν, τα οποία χρησιμοποιούνται για αναλύσεις τεραστίων βάσεων δεδομένων.

Ο τομέας της στατιστικής έχει να επιδείξει πολλά προγράμματα για αναλύσεις και παρουσιάσεις γραφημάτων, το κορυφαίο από αυτά θεωρείται το SPSS της IBM, μια πλατφόρμα επεξεργασίας λογιστικών φύλλων και γραφημάτων με δυνατότητες γραφικής απεικόνισης και ανάλυσης σε πραγματικό χρόνο. Ιδιαίτερα φιλικό προς τον χρήστη γραφικό περιβάλλον και απαιτώντας ελάχιστη εκπαίδευση στην χρήση του, το SPSS μπορεί να διαχειριστεί δεδομένα από βάσεις

δεδομένων σχεδόν κάθε τύπου, ακόμα και από διαφορετικά προγράμματα. Στα μειονεκτήματά του είναι η σχετικά αργή απόκρισή του σε μεγάλα λογιστικά φύλλα και η δυσκολία εισαγωγής δεδομένων από το Microsoft Excel λόγω διαφορετικού τρόπου επεξεργασίας.

Ένα άλλο ισχυρό εργαλείο επεξεργασίας στατιστικών και λογιστικών φύλλων είναι το Analytica της αμερικανικής Lumina. Χρησιμοποιεί την αρχιτεκτονική των απλουστευμένων φύλλων σε ένα δυναμικό περιβάλλον χωρίς να απαιτεί από τον χρήστη να πληκτρολογήσει κώδικα και του επιτρέπει να δημιουργεί πολυδιάστατες βάσεις δεδομένων και να εκτελεί σύνθετα ερωτήματα αλλά και να υπολογίζει τις πιο σημαντικές παραμέτρους των στατιστικών του και να τα παρακολουθεί.

Ο κλάδος της μηχανικής μάθησης τώρα, δεν στερείται την δική του γκάμα προγραμμάτων που μπορούν να εκπαιδεύσουν αποτελεσματικά ένα σύστημα έτσι ώστε να μπορέσει να αναγνωρίσει μοτίβα. Το γνωστότερο πρόγραμμα μηχανικής μάθησης είναι το Matlab, το οποίο χρησιμοποιεί πίνακες, στα δεδομένα των οποίων μπορούν να εκτελεστούν ταχύτατα και μπορεί να παρουσιάσει τα αποτελέσματά του σε γραφικό περιβάλλον. Στα μειονεκτήματα του η υψηλή τιμή του αλλά και η απαιτητική γλώσσα προγραμματισμού που χρησιμοποιεί.

Μια εναλλακτική χρήση είναι το πρόγραμμα WEKA, μια συλλογή αναλυτικών αλγορίθμων γραμμένη σε Java η οποία υποστηρίζει την συγγραφή παραμετροποιημένων προγραμμάτων με την αρχιτεκτονική της, λύνοντας έτσι προβλήματα που απαιτούν σύνθετη είσοδο από τον χρήστη. Το WEKA διατίθεται δωρεάν και η κοινότητα του γράφει συνεχώς νέους αλγόριθμους ανάλυσης και τους διαθέτει στο φόρουμ της.

Η τεχνητή νοημοσύνη αν και είναι πλέον κομμάτι της επιστήμης της εξόρυξης γνώσης έχει ελάχιστα εμπορικά προγράμματα, κυρίως λόγω του υψηλού ακόμη κόστους ανάπτυξης μιας αποδεκτού επιπέδου ανταπόκρισης στην ανθρώπινη είσοδο. Εντούτοις, η SAS προωθεί την ομώνυμη σουίτα προϊόντων παραγωγής στατιστικών, τα οποία χρησιμοποιούν έξυπνους αλγορίθμους οι οποίοι επεξεργάζονται και αναλύουν τις βάσεις δεδομένων του χρήστη σε πραγματικό χρόνο, μαθαίνοντας ταυτόχρονα τις συχνότερες απαιτήσεις του χρήστη. Η υψηλή τιμή του SAS και η ανάγκη για ένα υψηλών προδιαγραφών σύστημα είναι τα μειονεκτήματά που πιθανόν να απωθήσουν έναν πελάτη, αλλά πρόκειται για μια δυναμική πλατφόρμα που έχει κερδίσει δεκάδες βραβεία για την αποδοτικότητά της.

Οι διάφορες εμπορικές εφαρμογές των τομέων της εξόρυξης γνώσης έχουν ως στόχο την εξάπλωση της χρήσης της όχι μόνο σε υψηλού επιπέδου προγραμματιστικά περιβάλλοντα αλλά και τον μέσο χρήστη ο οποίος θέλει να κατανοήσει την επιστήμη της στατιστικής, να εκπαιδεύσει το δικό του σύστημα ή να πειραματιστεί με την τεχνητή νοημοσύνη. Προγράμματα δωρεάν αλλά και επί πληρωμή υπάρχουν για κάθε χρήση. Οι απαιτήσεις τους ποικίλουν, αλλά τόσο μια πολυεθνική επιχείρηση που έχει ανάγκη για επεξεργασία στατιστικών πινάκων με σκοπό τις μελλοντικές της κινήσεις όσο και ο απλός φοιτητής που χρησιμοποιεί ένα σύστημα μηχανικής μάθησης για να μάθει περισσότερα για την επίδοση της αθλητικής του ομάδας μπορούν να χρησιμοποιούν λογισμικό εξόρυξης γνώσης για την λήψη αυτών των αποτελεσμάτων ακόμα και αν δεν διαθέτουν πολλές φορές το κατάλληλο προγραμματιστικό υπόβαθρο.

Επίλογος

Σε αυτό το κεφάλαιο εξετάστηκαν οι κυριότερες εφαρμογές της εξόρυξης γνώσης, τόσο στο μάρκετινγκ και στις αγορές γενικότερα όσο και ασυνήθιστες εφαρμογές όπως η συνέργεια της εξόρυξης γνώσης στην διασφάλιση των διαδικτυακών συναλλαγών. Ακόμη, παρουσιάστηκαν ορισμένες από τις γνωστότερες εμπορικές εφαρμογές που η αγορά χρησιμοποιεί για να εξάγει συμπεράσματα από βάσεις δεδομένων.

Εν κατακλείδι, το συμπέρασμα είναι ότι σήμερα, με φθηνούς και ισχυρούς υπολογιστές οι χρήσεις της εξόρυξης γνώσης δεν περιορίζονται μόνο στην απλή στατιστική, αλλά παρεμβαίνουν δυναμικά, προβλέποντας μελλοντικά αποτελέσματα και καθορίζοντας τις κινήσεις εμπορικών κολοσσών. Τελικά, υπάρχει μια εφαρμογή και μία απάντηση για κάθε πρόβλημα. Στο επόμενο κεφάλαιο θα διεξαχθεί μια έρευνα με σκοπό την εξέταση των γνώσεων του μέσου χρήστη για την επιστήμη της εξόρυξης γνώσης.

Κεφάλαιο 5: Ερωτηματολόγιο

Εισαγωγή

Για να γίνει πιο κατανοητή η επιρροή της εξόρυξης γνώσης στην καθημερινότητα, δημιουργήθηκε αυτό το ερωτηματολόγιο που εξετάζει κατά πόσο είναι ενήμερος ο μέσος χρήστης τόσο για τις βασικές έννοιες της εξόρυξης γνώσης αλλά και για το ζήτημα της ασφάλειας των προσωπικών του δεδομένων. Το ερωτηματολόγιο αποτελείται από δεκαπέντε ερωτήσεις, σε τρεις ξεχωριστές ενότητες.

Η πρώτη αναζητά απαντήσεις σχετικές με τις γνώσεις του χρήστη σχετικά με την εξόρυξη γνώσης και τις πληροφορίες που μοιράζεται. Η δεύτερη ενότητα ασχολείται με την αγοραστική ψυχολογία του χρήστη και εξετάζει κατά πόσο επηρεάζεται από διαφημίσεις και cookies. Τέλος, η τρίτη θέτει το ερώτημα του κατά πόσο ο χρήστης εμπιστεύεται τους κατόχους των προσωπικών του δεδομένων όπως το κράτος, οι τράπεζες και το διαδίκτυο.

Το ερωτηματολόγιο μεταφράστηκε επίσης στα Αγγλικά και ανέβηκε σε γνωστή σελίδα πραγματοποίησης ερευνών, όπου συλλέχθηκαν περισσότερες απαντήσεις. Η ιστοσελίδα (surveymonkey) εγγυάται ότι κάθε απάντηση είναι μοναδική, αποκλείοντας τις διπλές εγγραφές. Επιπλέον, δεν απαιτείται λογαριασμός για να συμμετάσχει κάποιος στην έρευνα, και πουθενά δεν αναφέρονται ούτε ζητούνται τα προσωπικά στοιχεία του ερωτώμενου. Ακόμα, το ερωτηματολόγιο τυπώθηκε και μοιράστηκε σε όσο το δυνατόν μεγαλύτερη και πιο διαφοροποιημένη γκάμα ανθρώπων έτσι ώστε οι απαντήσεις να μην είναι μονόπλευρες. Ακολουθεί το ίδιο το ερωτηματολόγιο, ορισμένες χαρακτηριστικές απαντήσεις τόσο ηλεκτρονικής όσο και έντυπης μορφής και τα συμπεράσματα που καταγράφηκαν.

Ερωτηματολόγιο Εξόρυξη Γνώσης

Το ακόλουθο ερωτηματολόγιο χωρίζεται σε τρεις θεματικές ενότητες με πέντε ερωτήσεις στην καθεμία. Οι απαντήσεις κυμαίνονται στην κλίμακα 1 έως 5 με το 5 να δηλώνει απόλυτη συμφωνία και το 1 απόλυτη διαφωνία.

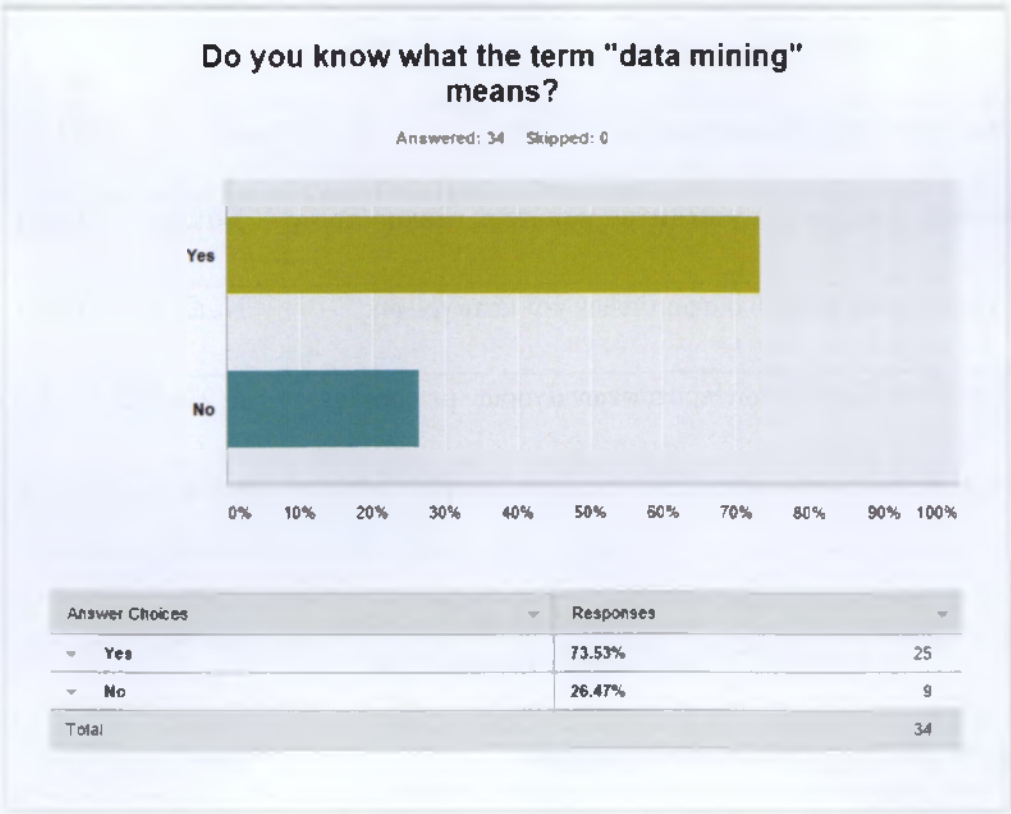
Ενότητα 1: Βασικές έννοιες	
1. Γνωρίζετε τι σημαίνει ο όρος «εξόρυξη γνώσης»;	Ναι❑ Όχι❑
2. Πιστεύετε ότι οι αγοραστικές επιλογές σας έχουν αντίκτυπο στην αγορά;	Ναι❑ Όχι❑
3. Γνωρίζετε πόσο μεγάλο είναι το ψηφιακό σας αποτύπωμα;	Ναι❑ Όχι❑
4. Θεωρείτε αναγκαία την χρήση διαφημίσεων στο διαδίκτυο;	1❑ 2❑ 3❑ 4❑ 5❑
5. Στα μέσα κοινωνικής δικτύωσης, μοιράζεστε προσωπικά δεδομένα;	1❑ 2❑ 3❑ 4❑ 5❑
Ενότητα 2: Εξόρυξη Γνώσης και Αγορές	
1. Συμμετέχετε σε κάποιο πρόγραμμα ανταποδοτικότητας;	Ναι❑ Όχι❑
2. Θα δίνατε προσωπικά στοιχεία με αντάλλαγμα καλύτερη εξυπηρέτηση;	Ναι❑ Όχι❑
4. Χρησιμοποιείτε μέσα αποκλεισμού διαφημίσεων και καταγραφής;	Ναι❑ Όχι❑
4. Είναι σημαντική η εξατομίκευση των προτάσεων αγορών με cookies;	1❑ 2❑ 3❑ 4❑ 5❑
5. Οι αγορές σας επηρεάζονται από διαφημίσεις που σας προβάλλονται:	1❑ 2❑ 3❑ 4❑ 5❑
Ενότητα 3: Ιδιωτικότητα και ασφάλεια	
1. Έχετε πέσει θύμα υποκλοπής προσωπικών στοιχείων;	Ναι❑ Όχι❑

2. Είναι σωστό να ερευνά το κράτος τα στοιχεία μας για λόγους ασφαλείας	Ναι❑ Όχι❑
3. Εμπιστεύεστε την τράπεζά σας με τα προσωπικά σας δεδομένα;	Ναι❑ Όχι❑
4. Πόσο σημαντική θεωρείτε την διαφύλαξη του απορρήτου;	1❑ 2❑ 3❑ 4❑ 5❑
5. Πόσο ασφαλή πιστεύετε ότι είναι τα στοιχεία σας στο διαδίκτυο;	1❑ 2❑ 3❑ 4❑ 5❑

Αποτελέσματα

Παρακάτω παρατίθενται τα αποτελέσματα τόσο της διαδικτυακής έρευνας όσο και των χειρόγραφων ερωτηματολογίων. Για λόγους οικονομίας χώρου τα χειρόγραφα έχουν ενσωματωθεί στα τελικά αποτελέσματα.

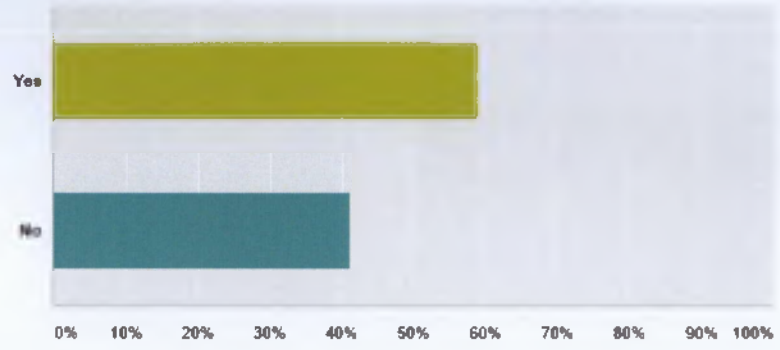
Διαδικτυακή Έρευνα



Answer Choices	Responses	
Yes	73.53%	25
No	26.47%	9
Total		34

Do you believe your buying choices affect the market as a whole?

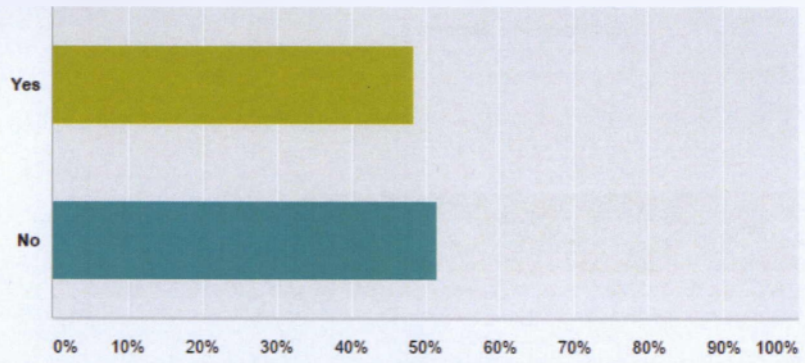
Answered: 34 Skipped: 0



Answer Choices	Responses	
Yes	58.82%	20
No	41.18%	14
Total		34

Are you aware of the size of your digital footprint?

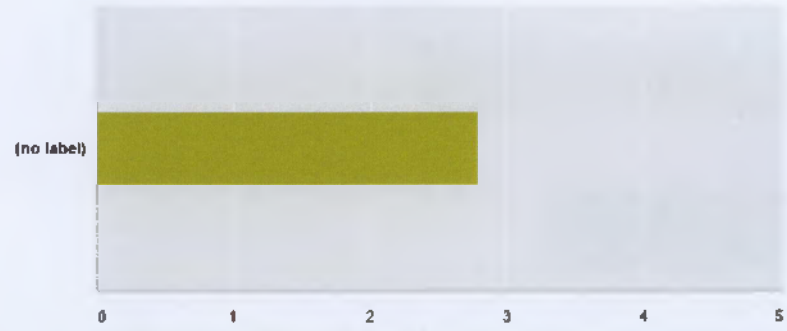
Answered: 33 Skipped: 1



Answer Choices	Responses	
Yes	48.48%	16
No	51.52%	17
Total		33

Do you share personal data on social networks you use?

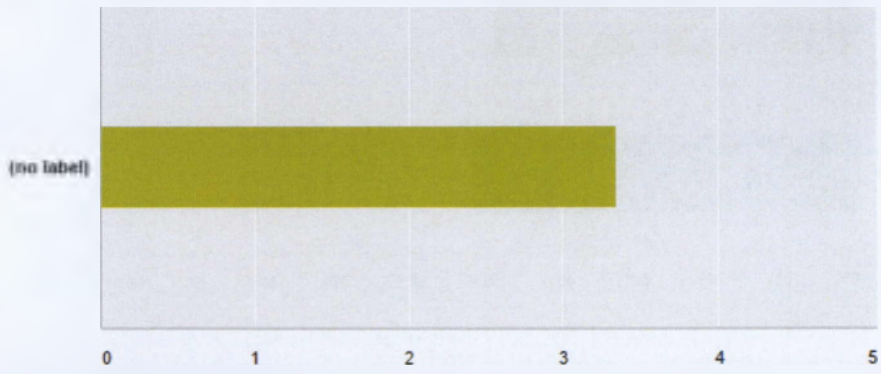
Answered: 34 Skipped: 0



	Almost no data, maximum privacy settings	Some data, only with friends	Fairly open profile, few areas locked	Open profile, personal information visible	I am actively sharing all my data with as many people as I can	I do not use social networks	Total	Average Rating
(no label)	5.88% 2	32.35% 11	26.47% 9	8.82% 3	8.82% 3	17.65% 6	34	2.79

Do you believe advertising on the internet is necessary?

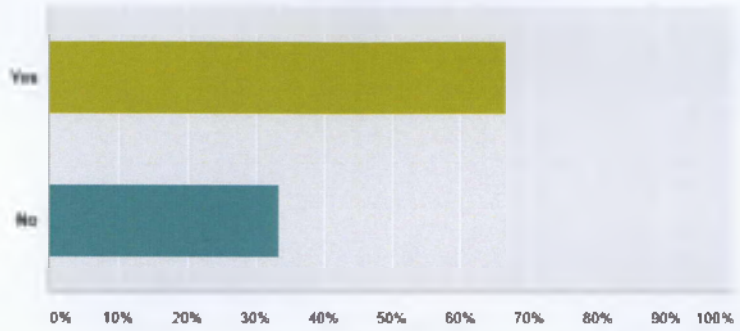
Answered: 34 Skipped: 0



	Strongly Disagree	Disagree	Neither Disagree Nor Agree	Agree	Strongly Agree	Total	Average Rating
(no label)	8.82% 3	11.76% 4	29.41% 10	35.29% 12	14.71% 5	34	3.35

Would you submit personal data in return for a better online shopping experience?

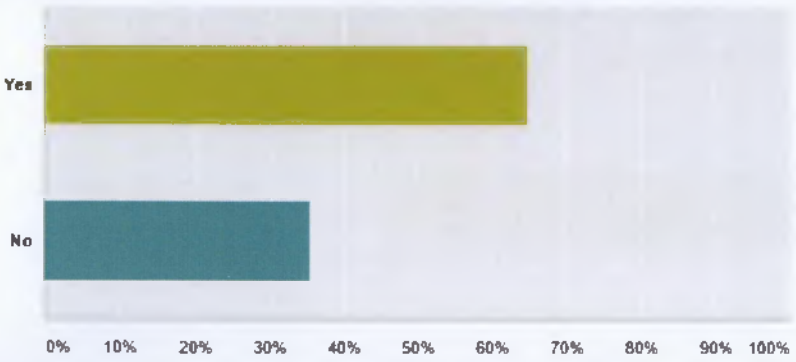
Answered: 33 Skipped: 1



Answer Choices	Responses	
Yes	66.67%	22
No	33.33%	11
Total		33

Are you currently using an advertising and cookies blocker while browsing? (AdBlock, Noscript, FlashBlock, DoNotTrack)

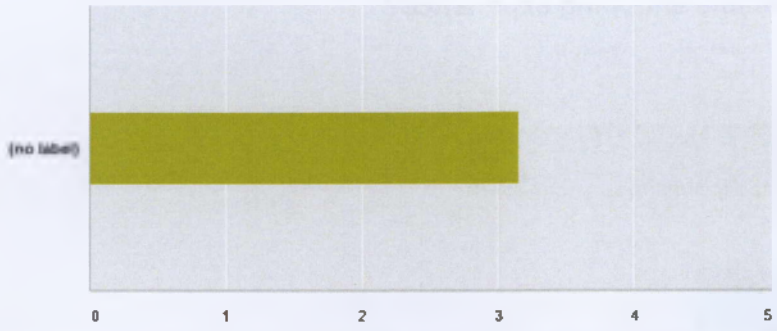
Answered: 34 Skipped: 0



Answer Choices	Responses	
Yes	64.71%	22
No	35.29%	12
Total		34

How affected are your purchases by ads you see?

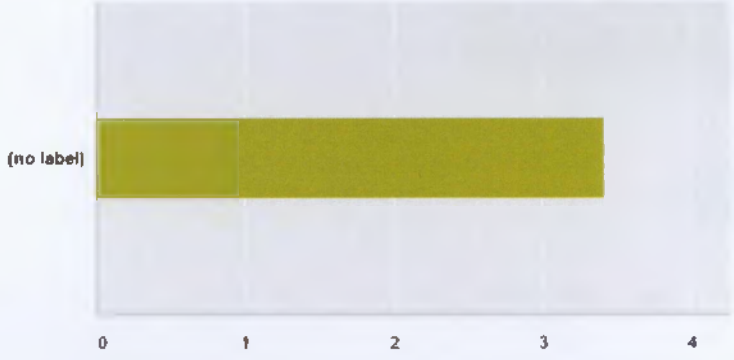
Answered: 34 Skipped: 0



	Totally Unaffected	Somewhat Unaffected	Neither Unaffected Nor Affected	Somewhat Affected	Totally Affected	Total	Average Rating
(no label)	11.76% 4	8.82% 3	35.29% 12	41.18% 14	2.94% 1	34	3.15

How important do you consider personalization of shopping services by using cookies and our personal data?

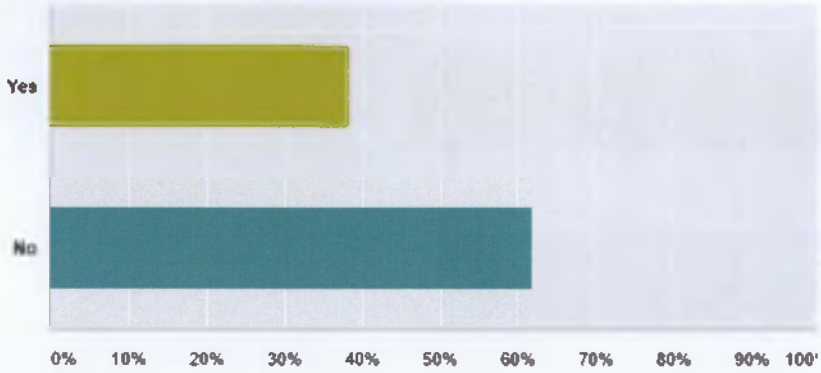
Answered: 34 Skipped: 0



	Strongly Disagree	Disagree	Neither Disagree Nor Agree	Agree	Strongly Agree	Total
(no label)	5.88% 2	11.76% 4	29.41% 10	41.18% 14	11.76% 4	34

Have you ever been a victim of data or identity theft?

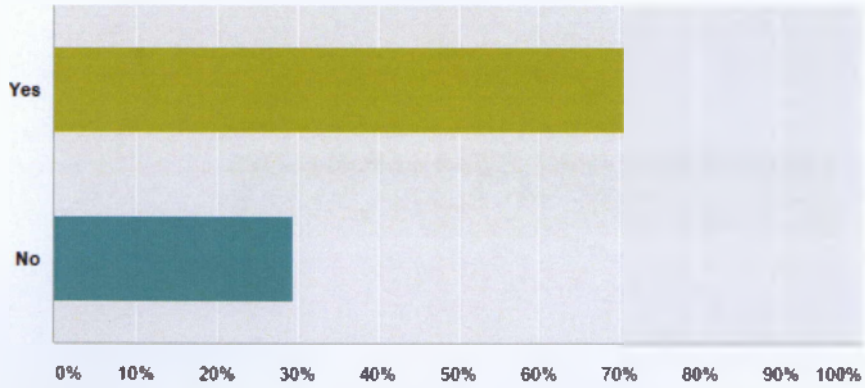
Answered: 34 Skipped: 0



Answer Choices	Responses	
Yes	38.24%	13
No	61.76%	21
Total		34

Should the State analyze and research our personal data for security reasons?

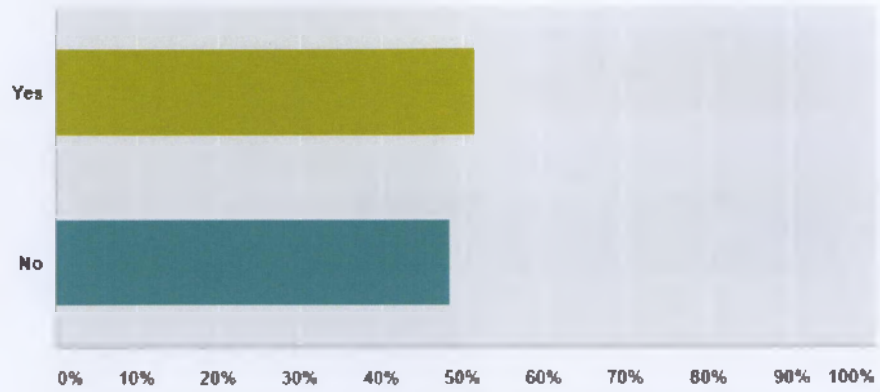
Answered: 34 Skipped: 0



Answer Choices	Responses	
Yes	70.59%	24
No	29.41%	10
Total		34

Do you trust your bank with your personal data?

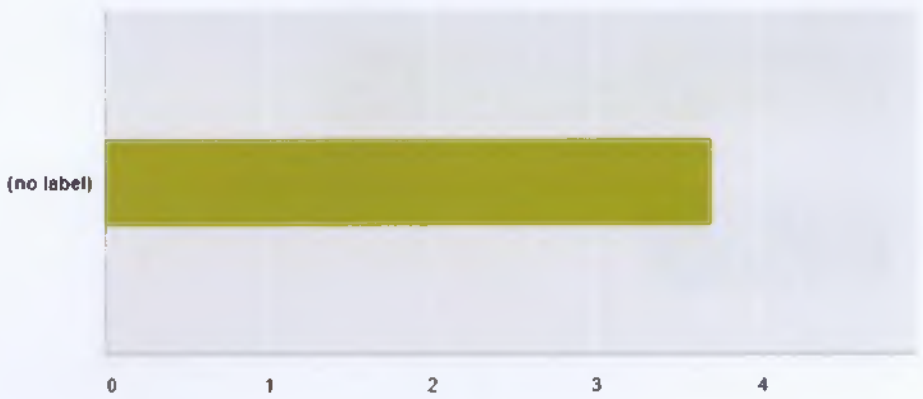
Answered: 33 Skipped: 1



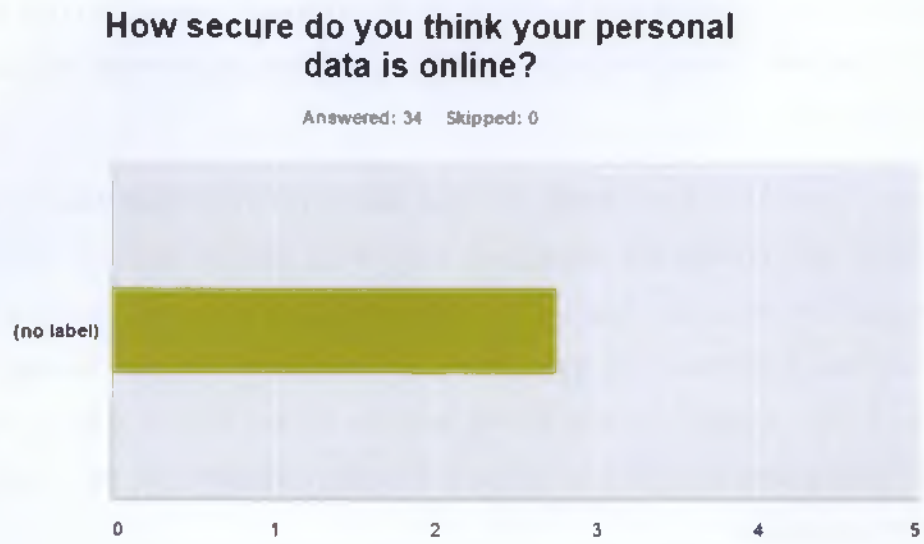
Answer Choices	Responses	
Yes	51.52%	17
No	48.48%	16
Total		33

How important is the assurance of privacy?

Answered: 34 Skipped: 0



	Extremely Unimportant	Not Important	Somewhat Important	Very Important	Extremely Important	Total
(no label)	0.00% 0	11.76% 4	26.47% 9	41.18% 14	20.59% 7	34



	Extremely Unsecure	Unsecure	Somewhat secure	Secure	Extremely Secure	Total	Average Rating
(no label)	5.88% 2	29.41% 10	47.06% 16	17.65% 6	0.00% 0	34	2.76

Συμπεράσματα

Η διαδικτυακή έρευνα κρίνεται επιτυχής και τα συμπεράσματά της πολύτιμα στην κατανόηση του σκοπού της, δηλαδή το κατά πόσο ο μέσος χρήστης κατανοεί την επιστήμη της εξόρυξης γνώσης και πως αντιλαμβάνεται την επιχειρηματική της πλευρά. Ακόμη, εξετάστηκε η ικανότητα της προστασίας των προσωπικών δεδομένων και το κατά πόσο η διαφήμιση επηρεάζει τις καταναλωτικές του συνήθειες.

Ξεκινώντας με την πρώτη ενότητα της έρευνας, τις βασικές έννοιες. Τα αποτελέσματα έδειξαν ότι ο μέσος χρήστης γνωρίζει τι σημαίνει ο όρος εξόρυξη γνώσης. Οι απαντήσεις σχετικά με το αν γνωρίζει ο ερωτώμενος το μέγεθος του ψηφιακού του αποτυπώματος, δηλαδή τα στοιχεία που

κατέχει το διαδίκτυο για το πρόσωπό του, καθώς και το πόσο στοιχεία μοιράζεται στα κοινωνικά δίκτυα έδειξαν ότι περίπου οι μισοί γνώριζαν, και δεν μοιράζονταν προσωπικά δεδομένα παρά μόνο με τους φίλους τους.

Συνεχίζοντας στην ενότητα της διαφήμισης, το κοινό έδειξε ότι είναι διατεθειμένο να δώσει προσωπικά στοιχεία με αντάλλαγμα καλύτερες υπηρεσίες, και οι απόψεις εμφανίζονται διχασμένες στην ερώτηση αν η διαφήμιση είναι απαραίτητη στο διαδίκτυο, πράγμα περίεργο εφόσον το 60% περίπου παραδέχεται ότι χρησιμοποιεί λογισμικό αποκλεισμού διαφημίσεων και cookies αλλά και το ότι οι επηρεάζεται από διαφημίσεις που βλέπει. Παρ' όλα αυτά, οι χρήστες συμφωνούν στην πλειοψηφία τους με την χρήση και αποδοχή cookies για την βελτίωση των αγοραστικών τους προτάσεων.

Ο τελευταίος τομέας ασχολήθηκε με την ιδιωτικότητα και την ασφάλεια των προσωπικών δεδομένων. Οι απαντήσεις εδώ κυμάνθηκαν σε αναμενόμενα πλαίσια, με τους ερωτώμενους να δηλώνουν ότι ενδιαφέρονται γι την ασφάλεια τους, θεωρώντας ότι τα προσωπικά τους στοιχεία δεν είναι ασφαλή στο διαδίκτυο. Αξίζει να σημειωθεί εδώ ότι μόνο οι μισοί περίπου εμπιστεύονται την τράπεζά τους με τα προσωπικά τους δεδομένα, αλλά πάνω από το 70% συμφωνούν στην χρήση τους από το κράτος για έρευνα με σκοπό την βελτίωση της ασφάλειας. Τέλος, το 60% δήλωσε ότι δεν έχει πέσει ποτέ θύμα ηλεκτρονικής και θεωρεί ότι η προστασία από επιθέσεις είναι ιδιαίτερα σημαντική.

Εν κατακλείδι, τα συμπεράσματα δείχνουν ότι ο μέσος χρήστης είναι πλέον ενημερωμένος για την χρήση των προσωπικών του δεδομένων με σκοπό την διαφήμιση, και δεν διατίθεται να τα δώσει σε κανέναν παρά μόνο σε διαπιστευμένους φορείς. Επίσης, οι διαφημίσεις τον επηρεάζουν, αλλά η ιδιωτικότητα του παραμένει σημαντική. Ακόμα, διαπιστώνεται ότι αν και σχεδόν κάθε σελίδα προσπαθεί να καταγράψει κάποια σχετικά δεδομένα με τον χρήστη που την περιηγείται, ο μέσος χρήστης προστατεύεται με προγράμματα αποκλεισμού διαφημίσεων. Το συμπέρασμα; Η εξόρυξη γνώσης είναι επιθυμητή γιατί βελτιώνει την ποιότητα της περιήγησης των χρηστών, αλλά οι χρήστες διστάζουν να μοιραστούν τα δεδομένα τους, αλλά η υπόσχεση για επιπλέον προνόμια και προσφορές ίσως τους δελεάσουν.

Γενικό Συμπέρασμα

Η εργασία είχε σκοπό την βαθύτερη κατανόηση κατ' αρχάς των τεχνικών και αλγορίθμων της εξόρυξης γνώσης. Μετά από την ανάλυσή τους το συμπέρασμα είναι ότι υπάρχει ένας αλγόριθμος για κάθε πρόβλημα αναζήτησης και κατανόησης δεδομένων, και διαφορετικοί τρόποι αντιμετώπισης ταιριάζουν σε διαφορετικές ανάγκες των χρηστών.

Επίσης, καταγράφηκαν οι γνωστότερες τεχνικές συλλογής δεδομένων που οι μεγάλες εταιρείες του διαδικτύου αλλά και του επιχειρηματικού κόσμου γενικότερα χρησιμοποιούν για να κατανοήσουν τις αγοραστικές τάσεις των πελατών τους. Οι τεχνικές αυτές χωρίστηκαν σε αυτές που προϋπήρχαν των υπολογιστών και εκείνες που αναπτύχθηκαν κάνοντας χρήση του διαδικτύου και των νέων τεχνολογιών.

Η έρευνα επεκτάθηκε στους τόπους που εφαρμόζεται, ανακαλύπτοντας τις ξεχωριστές προκλήσεις που θέτει π.χ. ο στρατός ή οι τράπεζες στην επιστήμη της εξόρυξης γνώσης και τι απαντήσεις υπάρχουν. Τα συμπεράσματα που προκύπτουν είναι ότι δεδομένα μπορούν να αναλυθούν από αυτούς τους φορείς για τις δικές τους μοναδικές ανάγκες χρησιμοποιώντας τα κατάλληλα προγράμματα και τα αποτελέσματα είναι συνήθως ικανοποιητικά.

Επίσης, εξετάστηκαν οι κυριότερες εφαρμογές της εξόρυξης γνώσης στην σύγχρονη αγορά. Από την δημιουργία εξατομικευμένων αγοραστικών προφίλ έως την στοχευόμενη διαφήμιση υπάρχει πλήθος εργαλείων που το μάρκετινγκ χρησιμοποιεί για να πετύχει το βέλτιστο αποτέλεσμα. Καταγράφηκαν επίσης τα κορυφαία προγράμματα ανάλυσης σε κάθε τομέα της εξόρυξης γνώσης (στατιστική, μηχανική μάθηση, τεχνητή νοημοσύνη) και ζυγίστηκαν τα πλεονεκτήματα και τα μειονεκτήματά τους.

Τέλος, τέθηκε ένα ερωτηματολόγιο σε μια μερίδα χρηστών τόσο σε έντυπη όσο και σε ηλεκτρονική μορφή, και οι απαντήσεις έδειξαν ότι ο μέσος χρήστης νοιάζεται για την ιδιωτικότητα του, αλλά είναι πρόθυμος να δώσει τα στοιχεία του με αντάλλαγμα καλύτερες διαδικτυακές υπηρεσίες. Επιπλέον, η πλειοψηφία είναι πλέον ενημερωμένη για το πόσο ασφαλή είναι τα δεδομένα τους στο διαδίκτυο και χρησιμοποιεί προγράμματα προστασίας.

Εν κατακλείδι, η εργασία αυτή πέτυχε τους στόχους που έθεσε μέσα από την έρευνα και την ανάλυση διαφόρων αξιόπιστων πηγών, και θα αποτελέσει το εφαλτήριο για την εισαγωγή του

συγγραφέα στον κόσμο της ανάλυσης και εξαγωγής συμπερασμάτων από το χάος που λέγεται διαδίκτυο.

Πηγές/Βιβλιογραφία

Βιβλία

Ahlemyer-Stubbe, A. A Practical Guide to Data Mining for Business and Industry

Baesens, B. Analytics in a Big Data World: The Essential Guide to Data Science and its Applications

Berry, M., Linoff, G. Data Mining Techniques

Cios, K., Pedrycz, W., Swiniarski, R., Kurgan, L. Data Mining, A Knowledge Discovery Approach

Coleman, S., Stubbe, A. A Practical Guide to Data Mining for Business and Industry

Dunham, M. Data Mining Techniques and Algorithms

Han, J., Kamber, M. Data Mining Concepts and Techniques

Kolb, J. Business Intelligence in Plain Language

Kuonen, D. A Statistical Perspective of Data Mining. Άρθρο. CRM Zine V.48

Larose, D. Discovering Knowledge in Data: An introduction to Data Mining

Mehra, R., Grover, L. The Lure of Statistics in Data Mining. Άρθρο. Journal of Statistics V. 16

Myatt, G. Making Sense of Data: A Practical Guide to Exploratory Data Analysis

Thornton, J. Truth from Trash

Tsiptsis, K., Chorianopoulos, A. Data Mining Techniques in CRM: Inside Customer Segmentation

Yuan, S., Abidin, A., Sloan, M., Wang J. Internet Advertising: An Interplay among Advertisers, Online Publishers, Ad Exchanges and Web Users

Zacharski, R. A Programmer's Guide to Data Mining

Zaki, M., Meira W. Data Mining and Analysis

Διατριβές

Flavian, C., Guinaliu, M. Consumer trust, perceived security and privacy policy. Διδακτορική Διατριβή. University of Zaragosa, Zaragosa, Spain

Hearst, M. Untangling text data mining. Πτυχιακή Εργασία. University of California, Berkley

Wu, X., Kumar, V. Top 10 algorithms for data mining. Έρευνα. University of Queensland, Brisbane, Australia

Berkhin, P. Survey of Clustering Data Mining Techniques. Διατριβή. Accrue Software Inc.

Friedman, J. Data Mining and Statistics: What's the Connection; Έρευνα. Πανεπιστήμιο Stanford, California

Schapire, R. Explaining Adaboost. Έρευνα. Πανεπιστήμιο Princeton

Εικόνες και Γραφήματα

1.1 http://perception.csl.illinois.edu/gpca/introduction/clustering2.gif

1.2 http://www.merl.com/publications/images/TR2008-065.png

1.3
http://www.hypertextbookshop.com/dataminingbook/public_version/artifacts/images/pictures/chpt2interSec3Fig1.jpg

1.4https://upload.wikimedia.org/wikipedia/commons/thumb/f/fb/PageRanks-Example.svg/2000px-PageRanks-Example.svg.png

1.5 http://commonsenseatheism.com/wp-content/uploads/2011/08/bayes-rule.png

2.1.1 http://www.mdpi.com/1660-4601/7/2/596/ag

2.1.2 http://promotionmagazine.promo.it/wp/wp-content/uploads/2013/09/ziliani.jpg

2.1.3
http://www.personal.kent.edu/~rmuhamma/GraphTheory/MyGraphTheory/Diagrams/g67.gif

2.1.4 http://www.heatonresearch.com/images/article/fun/class-0.png

2.2.1 http://iconicbiztech.com/images/category/357729e4f0119443b5b882ead1268677.jpg

2.2.2
https://m1.behance.net/rendition/modules/56035779/disp/33a9f4efaf4a2454ff6d0c2ee7cc1482.jp

2.2.3 http://www.codeproject.com/KB/IP/Crawler/WebCrawlerArchitecture.png

2.2.4 http://www.pendriveapps.com/wp-content/uploads/WebCookieSniffer.png

4.1 http://www.outotec.com/Global/Products%20and%20services/Software/HSC/HSC_data.jpg

4.2 http://licollider.files.wordpress.com/2012/02/screen-shot-2012-02-21-at-9-29-26-am.png

Διαδικτυακές Σελίδες και Άρθρα

Κυριότερα Προγράμματα Στατιστικής Ανάλυσης: http://www.capterra.com/statistical-analysis-software/

Ιστορική Διαδρομή της Εξόρυξης Γνώσης: http://www.sqldatamining.com/index.php/data-mining-basics/history-of-data-mining

Η λογική του Apriori: http://nikhilvithlani.blogspot.gr/2012/03/apriori-algorithm-for-data-mining-made.html

Ο αλγόριθμος Pagerank: http://www.slideshare.net/maimustafa566/page-rank-algorithm-33212250

Μπαεσιανή Λογική: http://research.cs.queensu.ca/home/xiao/dm.html

Υποδείγματα Στατιστικής Έρευνας: http://www.autonlab.org/tutorials/

Υπόδειγμα λειτουργίας Μπαεσιανής Λογικής: http://www.saedsayad.com/naive_bayesian.htm

Ανάλυση των Καρτών Μέλους: http://analyticsindiamag.com/mining-customer-loyalty-data-for-insights-on-purchasing-behaviour/

Εμπορικές εφαρμογές των Καρτών Μέλους: http://www.marketline.com/blog/loyalty-programs-utilizing-big-data-analytics-to-improve-crm-for-retailers/

Εξηγώντας τα Δεντροδιαγράμματα: http://support.sas.com/publishing/pubcat/chaps/57587.pdf

Συλλέγοντος Δεδομένα με Web Crawlers: https://www.udemy.com/blog/web-data-mining/

Το Διαδίκτυο και οι Crawlers: http://cacm.acm.org/blogs/blog-cacm/153780-data-mining-the-web-via-crawling/fulltext

Cookies, Πως οι Εταιρείες Ανακαλύπτουν Προσωπικά Δεδομένα στον Παγκόσμιο Ιστό: http://content.time.com/time/magazine/article/0,9171,2058205,00.html