1

Εξόρυξη Δεδομένων

Σε αυτό το εισαγωγικό κεφάλαιο αρχίζουμε με την πεμπτουσία της Εξόρυξης Δεδομένων και ιδιαίτερα για την αντιμετώπιση της Εξόρυξης Δεδομένων από διάφορους επιστημονικούς κλάδους που συνεισφέρουν στο πεδίο. Καλύπτουμε την "Αρχή του Bonferroni", η οποία είναι μία πραγματική προειδοποίηση σχετικά με την κατάχρηση της δυνατότητας να εξορύσσουμε δεδομένα. Επίσης το κεφάλαιο αυτό είναι το σημείο όπου συνοψίζουμε μερικές ιδέες που δεν ανήκουν στο πεδίο αυτό αλλά είναι χρήσιμες για την κατανόηση μερικών σημαντικών εννοιών της εξόρυξης. Οι ιδέες αυτές συμπεριλαμβάνουν το μέτρο TF.IDF για τη σπουδαιότητα μίας λέξης, τη συμπεριφορά των συναρτήσεων κατακερματισμού (hashing) και των καταλόγων (index), καθώς και ταυτότητες που περιέχουν το e, τη βάση των φυσικών λογαρίθμων. Τέλος, δίνουμε ένα περίγραμμα των θεμάτων που θα καλυφθούν στα πλαίσια του βιβλίου.

1.1 Τί είναι Εξόρυξη Δεδομένων;

Ο πλέον ευρέως αποδεκτός ορισμός της "Εξόρυξης Δεδομένων" είναι η ανακάλυψη των "μοντέλων" των δεδομένων. Ωστόσο, ένα "μοντέλο" μπορεί να είναι πολλών ειδών. Στη συνέχεια αναφέρουμε τις σημαντικότερες κατευθύνσεις στη μοντελοποίηση αυτή.

1.1.1 Στατιστική Μοντελοποίηση

Οι στατιστικολόγοι ήταν οι πρώτοι που χρησιμοποίησαν τον όρο "Εξόρυξη Δεδομένων". Αρχικά, "Εξόρυξη Δεδομένων" ή "αποκάλυψη δεδομένων" ήταν ένας μειωτικός όρος που αναφερόταν σε προσπάθειες να εξαχθεί πληροφορία που δεν υποστηρίζονταν από τα δεδομένα. Η Ενότητα 1.2 παρουσιάζει το είδος των λαθών που κάποιος μπορεί να υποπέσει προσπαθώντας να εξαγάγει κάτι που δεν βρίσκεται πράγματι μέσα στα δεδομένα. Τώρα, οι στατιστικολόγοι θεωρούν την Εξόρυξη Δεδομένων ως την κατασκευή ενός *στατιστικού* μοντέλου, δηλαδή, μία υποκείμενης κατανομής από όπου εξάγονται τα ορατά δεδομένα.

Παράδειγμα 1.1: Έστω ότι τα δεδομένα μας είναι ένα σύνολο αριθμών. Τα δεδομένα αυτά είναι πολύ απλούστερα από αυτά που θα μπορούσαν να εξορυχθούν, αλλά χρησιμεύουν ως ένα απλό παράδειγμα. Ένας στατιστικολόγος θα μπορούσε να αποφασίσει ότι τα δεδομένα ακολουθούν μία Γκαουσσιανή κατανομή και να χρησιμοποιήσει έναν τύπο για να υπολογίσει τις πιθανότερες παραμέτρους της Γκαουσσιανής κατανομής. Η μέση τιμή και η τυπική απόκλιση αυτής της Γκαουσσιανής κατανομής χαρακτηρίζουν πλήρως την κατανομή και θα μπορούσαν να αποτελέσουν το μοντέλο των δεδομένων. □

1.1.2 Μηχανική Μάθηση

Από αρκετούς η Εξόρυξη Δεδομένων θεωρείται συνώνυμη της Μηχανικής Μάθησης. Χωρίς αμφιβολία η Εξόρυξη Δεδομένων χρησιμοποιεί με κατάλληλο τρόπο αλγορίθμους της Μηχανικής Μάθησης. Οι επαγγελματίες της Μηχανικής Μάθησης χρησιμοποιούν τα δεδομένα σαν ένα σύνολο εκπαίδευσης για να εκπαιδεύσουν έναν αλγόριθμο από τους πολλούς της Μηχανικής Μάθησης, όπως είναι τα Δίκτυα Bayes (Bayes nets), οι Μηχανές Διανυσμάτων Υποστήριξης (support-vector machines), τα Δένδρα Αποφάσεων (decision trees), τα Κρυμμένα Μοντέλα Markov (hidden Markov models), και πολλοί άλλοι.

Υπάρχουν καταστάσεις όπου έχει νόημα αυτή η χρήση των δεδομένων. Η τυπική περίπτωση όπου η Μηχανική Μάθηση αποτελεί μία καλή προσέγγιση είναι όταν δεν έχουμε καλή αντίληψη για το τι αναζητούμε μέσα στα δεδομένα. Για παράδειγμα, είναι μάλλον ασαφές τι ακριβώς είναι αυτό που περιέχουν οι ταινίες, το οποίο κάνει τους θεατές να τους αρέσουν ή όχι. Έτσι, στην απάντηση της πρόκλησης της Netflix για την επινόηση ενός αλγορίθμου πρόβλεψης των κατατάξεων των ταινιών από τους χρήστες με βάση ένα δείγμα προηγούμενων απαντήσεών τους, οι αλγόριθμοι Μηχανικής Μάθησης αποδείχθηκαν αρκετά επιτυχείς. Θα συζητήσουμε αυτόν τον τύπο αλγορίθμων στην Ενότητα 9.4.

Από την άλλη πλευρά, η Μηχανική Μάθηση δεν αποδείχθηκε επιτυχής σε καταστάσεις, όπου μπορούμε να περιγράψουμε το στόχο της εξόρυξης περισσότερο άμεσα. Πιο συγκεκριμένα, μία ενδιαφέρουσα περίπτωση είναι η προσπάθεια των WhizBang! Labs [1] να χρησιμοποιήσουν τη Μηχανική Μάθηση για να εντοπίσουν βιογραφικά επαγγελματιών στον Παγκόσμιο Ιστό. Ωστόσο, δεν μπόρεσαν να τα καταφέρουν καλύτερα από ότι αλγόριθμοι σχεδιασμένοι με το χέρι για την αναζήτηση μερικών προφανών λέξεων και φράσεων που εμφανίζονται σε ένα τυπικό βιογραφικό. Καθώς ο καθένας που έχει κοιτάξει ή έχει γράψει ένα βιογραφικό έχει μία αντίληψη του περιεχομένου ενός βιογραφικού, δεν υπήρχε κάποιο μυστήριο σχετικά με το τι καθιστά βιογραφικό μία σελίδα του Παγκόσμιου Ιστού. Συνεπώς, δεν υπήρξε κάποιο πλεονέκτημα της Μηχανικής Μάθησης σε σχέση με τον άμεσο σχεδιασμό ενός αλγορίθμου για την ανακάλυψη βιογραφικών.

1.1.3 Υπολογιστικές Προσεγγίσεις Μοντελοποίησης

Πιο πρόσφατα, οι επιστήμονες της Πληροφορικής θεώρησαν την Εξόρυξη Δεδομένων ως ένα αλγοριθμικό πρόβλημα. Στην περίπτωση αυτή, το μοντέλο των δεδομένων είναι απλώς η απάντηση σε ένα σύνθετο ερώτημα για τα δεδομένα αυτά. Για παράδειγμα, δοθέντος του συνόλου αριθμών του Παραδείγματος 1.1, θα μπορούσαμε να υπολογίσουμε τη μέση τιμή και την τυπική απόκλιση. Σημειώνεται ότι οι τιμές αυτές μπορεί να μην είναι οι παράμετροι της Γκαουσσιανής κατανομής που ταιριάζει καλύτερα στα δεδομένα, αν και με μεγάλη ασφάλεια βρίσκονται πολύ κοντά αν το μέγεθος του συνόλου είναι μεγάλο.

Υπάρχουν πολλές διαφορετικές προσεγγίσεις για τη μοντελοποίηση των δεδομένων. Ήδη σημειώσαμε τη δυνατότητα κατασκευής μίας στατιστικής διαδικασίας με την οποία τα δεδομένα θα μπορούσαν να έχουν δημιουργηθεί. Άλλες προσεγγίσεις για τη μοντελοποίηση μπορούν να περιγραφούν είτε ως:

1. Η σύνοψη των δεδομένων με σύντομο και προσεγγιστικό τρόπο, ή

2. Η εξαγωγή του πλέον εξέχοντος γνωρίσματος των δεδομένων αγνοώντας τα υπόλοιπα.

Θα εξερευνήσουμε τις δύο αυτές προσεγγίσεις στη συνέχεια.

[1] Αυτή η εταιρεία start-up προσπάθησε να χρησιμοποιήσει Μηχανική Μάθηση για την Εξόρυξη Δεδομένων μεγάλης κλίμακας και για το σκοπό αυτό προσέλαβε τους κορυφαίους επιστήμονες της περιοχής. Δυστυχώς, η εταιρεία δεν κατάφερε να επιζήσει.

1.1.4 Σύνοψη

Μία από τις πιο ενδιαφέρουσες μορφές σύνοψης είναι η ιδέα του PageRank που κατέστησε τη Google τόσο επιτυχημένη και την οποία θα καλύψουμε στο Κεφάλαιο 5. Σε αυτόν τον τύπο εξόρυξης από τον Παγκόσμιο Ιστό, ολόκληρη η σύνθετη δομή του Παγκόσμιου Ιστού συνοψίζεται με έναν απλό αριθμό για κάθε σελίδα. Αυτός ο αριθμός, το "PageRank" αυτής της σελίδας είναι (με αρκετή υπεραπλούστευση) η πιθανότητα να βρίσκεται στη σελίδα αυτή σε μία οποιαδήποτε χρονική στιγμή ένας τυχαίος περιπατητής στον αντίστοιχο γράφο. Η αξιοσημείωτη ιδιότητα της κατάταξης αυτής είναι ότι αντανακλά πολύ πιστά τη "σπουδαιότητα" της κάθε σελίδας – το βαθμό επιθυμίας κάποιου τυπικού αναζητητή η σελίδα αυτή να επιστραφεί ως απάντηση στο ερώτημά του.

Ένας άλλος σημαντικός τύπος σύνοψης, η Ομαδοποίηση (clustering), θα καλυφθεί στο Κεφάλαιο 7. Εδώ, τα δεδομένα θεωρούνται ως σημεία σε ένα πολυδιάστατο χώρο. Σημεία που είναι "κοντά" στο χώρο αυτό, ανατίθενται στην ίδια ομάδα. Οι ίδιες οι ομάδες παρουσιάζονται συνοπτικά, πιθανώς με το κεντροειδές της ομάδας και τη μέση απόσταση των σημείων της ομάδας από το κεντροειδές. Αυτές οι συνόψεις των ομάδων αποτελούν τη σύνοψη του συνόλου δεδομένων.

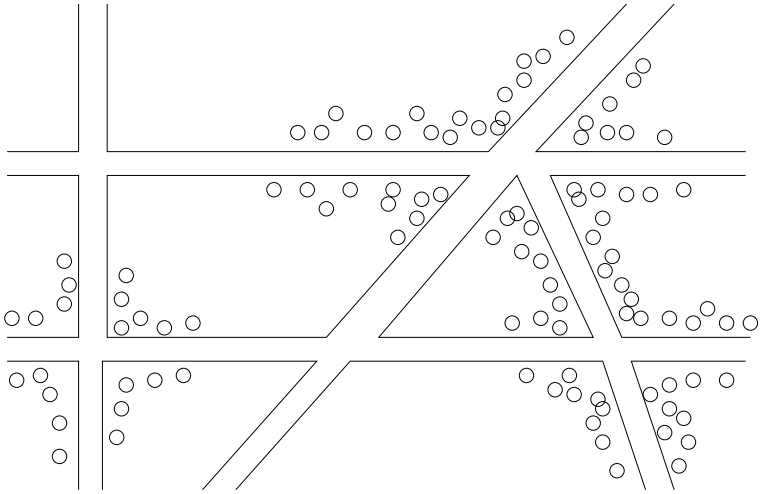
Παράδειγμα 1.2: Μία διάσημη περίπτωση Ομαδοποίησης για την επίλυση ενός προβλήματος έγινε προ πολλού στο Λονδίνο, και διεξήχθη εξ ολοκλήρου χωρίς υπολογιστές[2]. Ο γιατρός John Snow, που ασχολούνταν με το ξέσπασμα της χολέρας, σημείωσε τις θέσεις των περιστατικών σε ένα χάρτη της πόλης. Μία πρόχειρη εικονογράφηση που υποδηλώνει τη διαδικασία παρουσιάζεται στο Σχήμα 1.1.

Τα περιστατικά συγκεντρώνονταν γύρω από μερικές διασταυρώσεις δρόμων. Αυτές οι διασταυρώσεις ήταν οι τοποθεσίες πηγαδιών που είχαν μολυνθεί. Οι κάτοικοι που ζούσαν κοντά στα πηγάδια αυτά ασθένησαν, ενώ δεν ασθένησαν όσοι έμεναν κοντά σε μη μολυσμένα πηγάδια. Χωρίς τη δυνατότητα Ομαδοποίησης των δεδομένων, δεν θα είχε ανακαλυφθεί η αιτία της χολέρας. □

1.1.5 Εξαγωγή Γνωρισμάτων

Το τυπικό μοντέλο που βασίζεται στα γνωρίσματα αναζητεί τα πιο ακραία παραδείγματα ενός φαινομένου και αναπαριστά τα δεδομένα με αυτά τα παραδείγματα. Όποιος είναι εξοικειωμένος με τα Δίκτυα Bayes, έναν κλάδο Μηχανικής Μάθησης και ένα θέμα που δεν θα καλύψουμε στο βιβλίο αυτό, γνωρίζει πως αναπαρίσταται μία σύνθετη σχέση μεταξύ αντικειμένων βρίσκοντας τις ισχυρότε-

[2]Βλέπε `http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak`



Σχήμα 1.1: *Σημείωση των θέσεων των περιστατικών χολέρας σε ένα χάρτη του Λονδίνου.*

ρες στατιστικές εξαρτήσεις μεταξύ των αντικειμένων αυτών και χρησιμοποιώντας μόνο αυτές για την αναπαράσταση όλων των στατιστικών συσχετισμών. Μερικοί σημαντικοί τύποι εξαγωγής γνωρισμάτων από δεδομένα μεγάλης κλίμακας, τους οποίους θα μελετήσουμε είναι:

1. *Συχνά Είδη.* Αυτό το μοντέλο έχει νόημα για δεδομένα που αποτελούνται από "καλάθια" μικρών συνόλων ειδών, όπως στο πρόβλημα του "καλαθιού αγοράς", το οποίο θα συζητήσουμε στο Κεφάλαιο 6. Αναζητούμε μικρά σύνολα ειδών που εμφανίζονται μαζί σε πολλά καλάθια, και αυτά τα "συχνά είδη" είναι ο χαρακτηρισμός των δεδομένων που αναζητούμε. Η αρχική εφαρμογή αυτού του είδους εξόρυξης ήταν πραγματικά καλάθια αγορών: τα σύνολα των ειδών, όπως χάμπουργκερ και κέτσαπ, που οι πελάτες τείνουν να αγοράζουν μαζί, όταν πληρώνουν τον ταμία σε ένα κατάστημα ή ένα σούπερ μάρκετ.

2. *Παρόμοια Είδη.* Συχνά τα δεδομένα μας μοιάζουν με μία συλλογή συνόλων και ο σκοπός είναι να βρούμε ζεύγη συνόλων που έχουν κοινό ένα σχετικά μεγάλο ποσοστό ειδών. Ένα παράδειγμα είναι η αντιμετώπιση των πελατών ενός on-line καταστήματος, όπως η Amazon, σαν το σύνολο των ειδών που έχουν αγοράσει. Προκειμένου η Amazon να συστήσει κάτι διαφορετικό που θα άρεσε στους πελάτες, μπορεί να αναζητήσει "παρόμοιους" πελάτες και να συστήσει κάτι που έχουν αγοράσει πολλοί από αυτούς. Αυτή η διαδικασία ονομάζεται "συνεργατική διήθηση". Αν οι πελάτες ήταν αποφασισμένοι, δηλαδή, αν αγόραζαν μόνο ένα είδος προϊόντων, τότε θα

μπορούσε η Ομαδοποίηση να είναι αποδοτική. Ωστόσο, επειδή οι πελάτες έχουν ενδιαφέροντα για διάφορα προϊόντα, είναι χρησιμότερο για κάθε πελάτη να βρεθεί ένας μικρός αριθμός άλλων πελατών με παρόμοιες προτιμήσεις και να αναπαρασταθούν τα δεδομένα με αυτούς τους συσχετισμούς. Συζητούμε το θέμα της ομοιότητας στο Κεφάλαιο 3.

1.2 Στατιστικά Όρια της Εξόρυξης Δεδομένων

Ένας κοινός τύπος προβλημάτων Εξόρυξης Δεδομένων περιλαμβάνει την ανακάλυψη ασυνήθιστων γεγονότων κρυμμένων μέσα σε ογκώδεις ποσότητες δεδομένων. Στο σημείο αυτό θα συζητήσουμε αυτό το πρόβλημα, συμπεριλαμβανομένης της "αρχής Bonferroni", μίας προειδοποίησης εναντίον της υπερενθουσιώδους χρήσης της Εξόρυξης Δεδομένων.

1.2.1 Συνολική Επίγνωση της Πληροφορίας

Το 2002 η κυβέρνηση του Bush στις ΗΠΑ πρότεινε ένα σχέδιο για την εξόρυξη όλων των δεδομένων που μπορούσε να συλλέξει, συμπεριλαμβανομένων αποδείξεων από πιστωτικές κάρτες, καταχωρίσεις από ξενοδοχεία, ταξιδιωτικών δεδομένων και πολλά άλλα είδη πληροφοριών για να ιχνηλατήσει τρομοκρατικές δραστηριότητες. Φυσικά αυτή η ιδέα προκάλεσε μεγάλη ανησυχία μεταξύ των υποστηρικτών της ιδιωτικότητας, και τελικώς το πρότζεκτ της *Συνολικής Επίγνωσης της Πληροφορίας*, το επονομαζόμενο TIA (Total Information Awareness), αποσύρθηκε από το Κογκρέσο, μολονότι είναι ασαφές αν στην πραγματικότητα το πρότζεκτ υπάρχει με άλλη ονομασία. Δεν είναι σκοπός αυτού του βιβλίου να συζητήσει το δύσκολο ζήτημα του συμβιβασμού μεταξύ ιδιωτικότητας και ασφάλειας. Ωστόσο, η προοπτική του TIA ή ενός παρόμοιου συστήματος δεν εγείρει τεχνικές ερωτήσεις ως προς την εφικτότητα και το ρεαλισμό των υποθέσεών του.

Η ανησυχία πολλών ήταν ότι αν κοιτάξεις τόσα πολλά δεδομένα προσπαθώντας να εντοπίσεις δραστηριότητες που φαίνονται σαν τρομοκρατικές συμπεριφορές, δεν θα βρεις και πολλές αθώες δραστηριότητες ή ακόμη και έκνομες δραστηριότητες που δεν είναι όμως τρομοκρατικές, και οι οποίες θα καταλήξουν σε επισκέψεις από την αστυνομία ή και σε κάτι περισσότερο από μία απλή επίσκεψη; Η απάντηση είναι ότι όλα εξαρτώνται από το πόσο στενά ορίζονται οι αναζητούμενες δραστηριότητες. Οι στατιστικολόγοι έχουν εξετάσει αυτό το πρόβλημα σε πολλές εκδοχές και καθιέρωσαν μία θεωρία που εισάγουμε στη συνέχεια.

1.2.2 Αρχή Bonferroni

Έστω ότι διαθέτετε ένα πλήθος δεδομένων και αναζητάτε περιστατικά συγκεκριμένου τύπου στα δεδομένα αυτά. Μπορείτε να προσδοκάτε να εμφανισθούν τέτοιου τύπου περιστατικά ακόμη και αν τα δεδομένα είναι τελείως τυχαία, ενώ το πλήθος τους θα αυξάνεται καθώς θα μεγαλώνει και ο όγκος των δεδομένων. Αυτά τα περιστατικά είναι "κάλπικα", με την έννοια ότι δεν έχουν κάποιο λόγο διαφορετικό από το ότι πάντοτε τα τυχαία δεδομένα έχουν μερικά ασυνήθιστα γνωρίσματα που μοιάζουν σημαντικά αλλά δεν είναι. Ένα θεώρημα της Στατιστικής, γνωστό ως η "Διόρθωση του Bonferroni", δίνει ένα ασφαλή στατιστικό τρόπο ώστε να αποφεύγουμε την πλειοψηφία αυτών των κάλπικων απαντήσεων κατά την αναζήτηση στα δεδομένα. Χωρίς να υπεισερχόμαστε σε στατιστικές λεπτομέρειες, παραθέτουμε μία άτυπη εκδοχή της "Αρχής Bonferroni", η οποία μας βοηθά να αποφεύγουμε την αντιμετώπιση των τυχαίων περιστατικών ως πραγματικών. Υπολογίζουμε την αναμενόμενη τιμή των ειδικών περιστατικών που αναζητούμε υποθέτοντας ότι τα δεδομένα είναι τυχαία. Αν αυτός ο αριθμός είναι σημαντικά μεγαλύτερος από τον αριθμό των πραγματικών περιστατικών που ελπίζουμε να βρούμε, τότε κατά πάσα πιθανότητα σχεδόν το σύνολο των ευρημάτων μας θα είναι κάλπικα, δηλαδή περισσότερο ένα στατιστικό κατασκεύασμα παρά μία ένδειξη αυτού που ψάχνουμε. Αυτή η παρατήρηση είναι μία άτυπη δήλωση της Αρχής Bonferroni.

Σε ένα ενδεχόμενο που αναζητούμε τρομοκράτες, όπου αναμένουμε ότι σε μία δεδομένη στιγμή ενεργούν πολύ λίγοι τρομοκράτες, η Αρχή του Bonferroni λέει ότι μπορούμε να ανιχνεύσουμε τρομοκράτες μόνο αν ψάχνουμε για γεγονότα που είναι τόσο σπάνια, ώστε να είναι αδύνατο να εμφανισθούν σε τυχαία δεδομένα. Στη συνέχεια θα δώσουμε ένα εκτενές παράδειγμα.

1.2.3 Παράδειγμα της Αρχής Bonferroni

Ας υποθέσουμε ότι υπάρχουν κάποιοι "κακοί" και θέλουμε να τους εντοπίσουμε. Έστω επίσης ότι έχουμε λόγους να πιστεύουμε ότι περιοδικά οι "κακοί" συγκεντρώνονται σε ένα ξενοδοχείο για να συνωμοτήσουν. Ας κάνουμε τις εξής υποθέσεις σχετικά με το μέγεθος του προβλήματος:

1. Υπάρχουν 1.000.000.000 άνθρωποι που θα μπορούσε να είναι "κακοί".
2. Ο καθένας πηγαίνει σε ένα ξενοδοχείο 1 μέρα στις 100.
3. Ένα ξενοδοχείο έχει χωρητικότητα 100 ανθρώπων. Άρα, υπάρχουν 100.000 ξενοδοχεία, δηλαδή αρκετά για να στεγάσουν το 1% του 1.000.000.000 ανθρώπων που επισκέπτονται ένα ξενοδοχείο σε μία δεδομένη ημέρα.

4. Θα εξετάσουμε τα στοιχεία των ξενοδοχείων για 1.000 ημέρες.

Για να βρούμε τους "κακούς" μέσα σε αυτά τα δεδομένα, θα ψάξουμε για ανθρώπους που σε διαφορετικές ημέρες βρίσκονταν στο ίδιο ξενοδοχείο. Ωστόσο, ας υποθέσουμε ότι στην πραγματικότητα δεν υπάρχουν "κακοί". Δηλαδή, ο καθένας συμπεριφέρεται τυχαία και αποφασίζει με πιθανότητα 0,01 να πάει σε ένα ξενοδοχείο σε μία συγκεκριμένη ημέρα, ενώ επιλεγεί τυχαία ένα ξενοδοχείο από τα 10^5. Μπορούμε να βρούμε κάποιο ζεύγος ανθρώπων που να εμφανίζονται σαν "κακοί";

Μπορούμε να κάνουμε έναν απλό προσεγγιστικό υπολογισμό ως εξής. Η πιθανότητα 2 οποιοιδήποτε άνθρωποι να αποφασίσουν το ίδιο και να επισκεφθούν ένα ξενοδοχείο την ίδια ημέρα είναι 0,0001. Η πιθανότητα αυτοί οι άνθρωποι να επισκεφθούν το ίδιο ξενοδοχείο είναι η προηγούμενη πιθανότητα δια του 10^5, όσο το πλήθος των ξενοδοχείων. Συνεπώς, η πιθανότητα να επισκεφθούν το ίδιο ξενοδοχείο σε μία συγκεκριμένη ημέρα είναι 10^{-9}. Η πιθανότητα ότι θα επισκεφθούν το ίδιο ξενοδοχείο σε 2 διαφορετικές ημέρες είναι το τετράγωνο του αριθμού αυτού, δηλαδή 10^{-18}. Σημειώνεται επίσης ότι τα ξενοδοχεία μπορεί να είναι διαφορετικά στις 2 αυτές ημέρες.

Τώρα πρέπει να θεωρήσουμε πόσα γεγονότα θα δείξουν "κακές" ενέργειες. Με την έννοια αυτή ένα "γεγονός" είναι ένα ζεύγος ανθρώπων και ένα ζεύγος ημερών, έτσι ώστε οι 2 άνθρωποι να επισκεφθούν το ίδιο ξενοδοχείο στις 2 αυτές ημέρες. Για να απλοποιήσουμε την αριθμητική, σημειώνουμε ότι για μεγάλα n, το $\binom{n}{2}$ προσεγγίζει το $n^2/2$. Στη συνέχεια θα χρησιμοποιήσουμε την προσέγγιση αυτή. Έτσι, το πλήθος των ζευγών ανθρώπων είναι $\binom{10^9}{2} = 5 \times 10^{17}$. Το πλήθος των ζευγών ημερών είναι $\binom{1000}{2} = 5 \times 10^5$. Η αναμενόμενη τιμή γεγονότων που μοιάζουν "κακά" είναι το γινόμενο του πλήθους των ζευγών ανθρώπων, του πλήθους των ζευγών ημερών και της πιθανότητας ένα οποιοδήποτε ζεύγος ανθρώπων και ένα οποιοδήποτε ζεύγος ημερών να είναι ένα περιστατικό της συμπεριφοράς που αναζητούμε. Ο αριθμός αυτός είναι:

$$5 \times 10^{17} \times 5 \times 10^5 \times 10^{-18} \;=\; 250.000$$

Συνεπώς, θα υπάρχουν 250.000 ζεύγη ανθρώπων που μοιάζουν "κακοί" ακόμη και αν δεν είναι.

Τώρα, ας υποθέσουμε ότι στην πραγματικότητα υπάρχουν 10 ζεύγη "κακών". Η αστυνομία θα χρειασθεί να διερευνήσει 250.000 ζεύγη για να βρει τους πραγματικά "κακούς". Εκτός από την εισβολή στη ζωή 500.000 αθώων ανθρώπων, η απαραίτητη δουλειά είναι αρκετά τεράστια έτσι ώστε να καθίσταται ανέφικτη αυτή η προσέγγιση ανεύρεσης "κακών".

1.2.4 Ασκήσεις Ενότητας 1.2

Άσκηση 1.2.1: Με βάση τις πληροφορίες της Ενότητας 1.2.3, ποιός είναι ο αριθμός των ύποπτων ζευγών αν τα δεδομένα αλλάξουν ως εξής (ενώ όλοι οι άλλοι αριθμοί παραμένουν οι ίδιοι);

(α) Το πλήθος των ημερών υπό παρατήρηση ανέρχεται σε 2.000.

(β) Το πλήθος των ανθρώπων υπό παρακολούθηση ανέρχεται σε 2.000.000 και συνεπώς υπάρχουν 200.000 ξενοδοχεία.

(γ) Θα αναφερθεί ένα ζεύγος ως ύποπτο αν συνυπήρξαν στο ίδιο ξενοδοχείο ταυτοχρόνως σε 3 διαφορετικές ημέρες.

Άσκηση 1.2.2 (!)**:** Ας υποθέσουμε ότι έχουμε πληροφορίες για τις αγορές 100.000.000 ανθρώπων από σούπερ-μάρκετ. Ο καθένας πηγαίνει 100 φορές το χρόνο στο σούπερ-μάρκετ και αγοράζει 10 από τα 1000 πωλούμενα είδη του σούπερ-μάρκετ. Πιστεύουμε ότι ένα ζεύγος τρομοκρατών θα αγοράσει ακριβώς τα ίδια 10 είδη (πιθανώς τα συστατικά για μία βόμβα;) σε κάποια χρονική στιγμή εντός του έτους. Αν αναζητούμε ζεύγη ανθρώπων που αγόρασαν το ίδιο σύνολο ειδών, θα αναμέναμε ότι οποιοσδήποτε τέτοιος ύποπτος θα ήταν πραγματικά τρομοκράτης;

1.3 Πράγματα χρήσιμα να γνωρίζουμε

Στο κεφάλαιο αυτό επιχειρούμε μία συνοπτική εισαγωγή σε αντικείμενα που μπορεί να είναι γνωστά ή όχι από τη μελέτη άλλων αντικειμένων. Το καθένα είναι χρήσιμο για τη μελέτη της Εξόρυξης Δεδομένων. Τα αντικείμενα αυτά συμπεριλαμβάνουν:

1. Το μέτρο TF.IDF για τη σπουδαιότητα μίας λέξης.
2. Τις συναρτήσεις κατακερματισμού και τη χρήση τους.
3. Τη δευτερεύουσα αποθήκευση (δίσκος) και τη συνέπειά της στο χρόνο εκτέλεσης των αλγορίθμων.
4. Τη βάση e των φυσικών λογαρίθμων και ταυτότητες που εμπεριέχουν τη σταθερά αυτή.
5. Τους νόμους δυνάμεων.

1.3.1 Σπουδαιότητα των Λέξεων σε ένα Έγγραφο

Σε αρκετές εφαρμογές της Εξόρυξης Δεδομένων, θα αντιμετωπίσουμε το πρόβλημα της Κατηγοριοποίησης εγγράφων (μίας ακολουθίας λέξεων) με βάση το θέμα τους. Συνήθως, τα θέματα αναγνωρίζονται βρίσκοντας κάποιες ειδικές λέξεις που χαρακτηρίζουν τα έγγραφα σχετικά με το κάθε θέμα. Για παράδειγμα, άρθρα σχετικά με το baseball θα έτειναν να έχουν περιπτώσεις λέξεων όπως "`ball`", "`bat`", "`pitch`", "`run`" κ.ο.κ. Από τη στιγμή που έχουμε κατηγοριοποιήσει κάποια έγγραφα ότι αναφέρονται στο baseball, δεν είναι δύσκολο να παρατηρήσουμε ότι λέξεις σαν αυτές συνήθως εμφανίζονται συχνά. Ωστόσο, μέχρι να ολοκληρώσουμε την Κατηγοριοποίηση, δεν είναι δυνατό να αναγνωρίσουμε τις λέξεις αυτές ως γνωρίσματα.

Έτσι, συνήθως η Κατηγοριοποίηση αρχίζει κοιτώντας σε έγγραφα και βρίσκοντας τις σημαντικές λέξεις σε αυτά τα έγγραφα. Η πρώτη μας εικασία θα μπορούσε να είναι ότι οι λέξεις που εμφανίζονται συχνότερα σε ένα έγγραφο είναι και οι σημαντικότερες. Ωστόσο, αυτή η διαίσθηση είναι ακριβώς αντίθετη από την αλήθεια. Με μεγαλύτερη βεβαιότητα, συχνότερες λέξεις θα είναι κοινές λέξεις όπως "`the`" ή "`and`", που βοηθούν να αναπτύξουμε ιδέες αλλά δεν φέρουν κάποια σημαντικότητα από μόνες τους. Στην πράξη, μερικές εκατοντάδες κοινές Αγγλικές λέξεις (που ονομάζονται *λέξεις στάσης*, (stop words)) συχνά αγνοούνται πριν από κάθε προσπάθεια Κατηγοριοποίησης των εγγράφων.

Στην πραγματικότητα, οι δείκτες κάθε θέματος είναι οι σχετικά σπάνιες λέξεις. Ωστόσο, δεν είναι όλες οι σπάνιες λέξεις εξ ίσου χρήσιμες για δείκτες. Υπάρχουν συγκεκριμένες λέξεις, όπως για παράδειγμα "`παρότι`" ή "`μολονότι`", που εμφανίζονται σπάνια σε μία συλλογή εγγράφων αλλά δεν δηλώνουν κάτι χρήσιμο. Από την άλλη πλευρά, μία λέξη όπως "`chukker`" είναι πιθανώς εξ ίσου σπάνια αλλά μας στρέφει στη διαπίστωση ότι το έγγραφο σχετίζεται με το άθλημα του polo[3]. Η διαφορά μεταξύ σπάνιων λέξεων που κάτι δηλώνουν και αυτών που δεν δηλώνουν κάτι σχετίζεται με τη συγκέντρωση των σπάνιων χρήσιμων λέξεων σε μερικά μόνο έγγραφα. Δηλαδή, η παρουσία της λέξης "`μολονότι`" σε ένα έγγραφο δεν την καθιστά πολύ πιθανότερη αν παρουσιασθεί περισσότερες φορές. Ωστόσο, αν ένα άρθρο περιέχει τη λέξη "`chukker`" μία φορά, τότε είναι πιθανό να αναφέρει τι συνέβη στο "`πρώτο chukker`", ύστερα στο "`δεύτερο chukker`" κ.ο.κ. Δηλαδή, αν η λέξη παρουσιασθεί, τότε είναι πιθανό να επαναληφθεί.

Το επίσημο μέτρο της συγκέντρωσης των εμφανίσεων μίας συγκεκριμένης λέξης σε σχετικά λίγα έγγραφα ονομάζεται TF.IDF (*Term Frequency* × *Inverse*

[3]Σ.τ.Μ. το polo παίζεται από ομάδες 4 έφιππων παικτών σε 6 επτάλεπτες περιόδους-chukkers με τρίλεπτα διαλείμματα και πεντάλεπτο ημίχρονο.

Document Frequency)[4]. Συνήθως υπολογίζεται ως εξής. Έστω μία συλλογή N εγγράφων. Ως f_{ij} ορίζεται η συχνότητα (δηλαδή, το πλήθος ενφανίσεων) του όρου (λέξης) i στο έγγραφο j. Έτσι, η *Συχνότητα Όρου* ορίζεται ως:

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

Με άλλα λόγια, η Συχνότητα Όρου του όρου i στο έγγραφο j είναι η f_{ij} κανονικοποιημένη διαιρώντας την με το μέγιστο πλήθος εμφανίσεων οποιουδήποτε όρου (ίσως εξαιρώντας τις λέξεις στάσης) στο ίδιο έγγραφο. Έτσι, ο συχνότερος όρος στο έγγραφο j δίνει TF ίσο με 1, ενώ οι υπόλοιποι όροι λαμβάνουν Συχνότητα Όρου μικρότερη του 1 για το έγγραφο αυτό.

Η ποσότητα IDF για έναν όρο ορίζεται ως εξής. Έστω ότι ο όρος i εμφανίζεται σε n_i από τα N έγγραφα της συλλογής. Στην περίπτωση αυτή ισχύει $IDF_i = \log_2(N/n_i)$. Το αποτέλεσμα του TF.IDF για τον όρο i στο έγγραφο j ορίζεται ως $TF_{ij} \times IDF_i$. Οι όροι με υψηλότερο αποτέλεσμα TF.IDF είναι συνήθως οι όροι που χαρακτηρίζουν καλύτερα το θέμα του εγγράφου.

Παράδειγμα 1.3: Ας υποθέσουμε ότι η συλλογή μας αποτελείται από $2^{20} = 1.048.576$ έγγραφα. Έστω επίσης ότι η λέξη w εμφανίζεται σε 2^{10}=1024 από αυτά τα έγγραφα. Τότε $IDF_w = \log_2(2^{20}/2^{10}) = 10$. Ας θεωρήσουμε ένα έγγραφο j όπου η λέξη w εμφανίζεται 20 φορές, το οποίο είναι και το μέγιστο πλήθος εμφάνισης οποιασδήποτε λέξης (ίσως μετά την εξάλειψη των λέξεων στάσης). Τότε TF_{wj}=1, ενώ το αποτέλεσμα του TF.IDF για τη λέξη w στο έγγραφο j είναι 10.

Ας υποθέσουμε ότι η λέξη w εμφανίζεται 1 φορά στο έγγραφο k, ενώ το μέγιστο πλήθος εμφάνισης οποιασδήποτε λέξης στο έγγραφο αυτό είναι 20. Τότε TF_{wk}=1/20, ενώ το αποτέλεσμα TF.IDF της λέξης w στο έγγραφο k είναι 1/2. □

1.3.2 Συναρτήσεις Κατακερματισμού

Πιθανώς ο αναγνώστης να έχει ακούσει για πίνακες κατακερματισμού και ίσως να τους χρησιμοποίησε σε κλάσεις της Java ή παρόμοια πακέτα. Οι συναρτήσεις κατακερματισμού που καθιστούν εφικτούς τους πίνακες κατακερματισμού είναι επίσης ουσιώδη συστατικά σε αρκετούς αλγορίθμους Εξόρυξης Δεδομένων, όπου οι πίνακες κατακερματισμού λαμβάνουν ένα ασυνήθιστο σχήμα. Στη συνέχεια θα επανεξετάσουμε τα βασικά.

[4]Σ.τ.Μ. σε μετάφραση "Συχνότητα Όρου επί Αντίστροφη Συχνότητα Εγγράφου"

Πρώτον, μία συνάρτηση κατακερματισμού h λαμβάνει ως όρισμα μία τιμή *κλειδιού κατακερματισμού* και παράγει ως αποτέλεσμα έναν *αριθμό κάδου*. Ο αριθμός κάδου είναι ένας ακέραιος στο εύρος $[0..B-1]$, όπου B είναι το πλήθος των κάδων. Τα κλειδιά κατακερματισμού μπορεί να είναι οποιουδήποτε τύπου. Υπάρχει μία διαισθητική ιδιότητα των συναρτήσεων κατακερματισμού ότι αυτές "τυχαιοποιούν" τα κλειδιά κατακερματισμού. Για την ακρίβεια, αν τα κλειδιά κατακερματισμού παράγονται τυχαία από ένα λογικό πληθυσμό σχετικών κλειδιών, τότε η συνάρτηση h θα στείλει σχεδόν τον ίδιο αριθμό κλειδιών σε κάθε έναν κάδο από τους B. Αυτό θα ήταν αδύνατο να επιτευχθεί αν, για παράδειγμα, ο πληθυσμός των κλειδιών κατακερματισμού ήταν μικρότερος από B. Αυτός ο πληθυσμός θα ήταν "μη λογικός". Ωστόσο, μπορεί να υπάρξουν περισσότερο σοβαροί λόγοι γιατί μία συνάρτηση κατακερματισμού αδυνατεί να επιτύχει μία σχεδόν ομοιόμορφη κατανομή στους κάδους.

Παράδειγμα 1.4: Ας υποθέσουμε ότι τα κλειδιά κατακερματισμού είναι θετικοί ακέραιοι. Μία κοινή και απλή συνάρτηση κατακερματισμού είναι $h(x) = x \bmod B$, δηλαδή το υπόλοιπο της διαίρεσης του x δια B. Η επιλογή αυτή λειτουργεί άψογα αν ο πληθυσμός των κλειδιών είναι όλοι θετικοί ακέραιοι. Το $(1/B)$-οστό των ακεραίων θα ανατεθούν σε κάθε κάδο. Ωστόσο, ας υποθέσουμε ότι ο πληθυσμός μας αποτελείται από άρτιους ακεραίους και ότι $B = 10$. Τότε, η τιμή της $h(x)$ μπορεί να είναι μόνο οι κάδοι 0, 2, 4, 6, 8, οπότε σαφώς η συνάρτηση κατακερματισμού δεν είναι τυχαία στη συμπεριφορά της. Από την άλλη πλευρά, αν επιλέγαμε B=11, τότε θα διαπιστώναμε ότι το $(1/11)$-οστό των θετικών ακεραίων θα κατευθύνονταν στον κάθε ένα από τους 11 κάδους, οπότε η συνάρτηση θα λειτουργούσε ικανοποιητικά. □

Η γενίκευση του Παραδείγματος 1.4 είναι ότι όταν τα κλειδιά κατακερματισμού είναι ακέραιοι, τότε η επιλογή του B έτσι ώστε να έχει κοινούς παράγοντες με όλα (ή με τα περισσότερα από) τα κλειδιά κατακερματισμού, θα οδηγήσει σε μη ομοιόμορφες κατανομές στους κάδους. Έτσι, κανονικά προτιμάται να επιλέγουμε έναν πρώτο αριθμό για τιμή του B. Η επιλογή αυτή μειώνει την πιθανότητα για μία μη τυχαία συμπεριφορά αν και ακόμη πρέπει να θεωρήσουμε την πιθανότητα όπου όλα τα κλειδιά κατακερματισμού έχουν το B ως παράγοντα. Προφανώς υπάρχουν πολλοί άλλοι τύποι συναρτήσεων κατακερματισμού, οι οποίοι δεν στηρίζονται στην αριθμητική της διαίρεσης. Δεν θα προσπαθήσουμε να συνοψίσουμε εδώ τις επιλογές αυτές, αλλά θα αναφέρουμε μερικές πληροφοριακές πηγές στις βιβλιογραφικές σημειώσεις.

Τι θα συμβεί αν τα κλειδιά κατακερματισμού δεν είναι ακέραιοι; Με μία έννοια, όλοι οι τύποι δεδομένων έχουν τιμές που αποτελούνται από bits, οπότε οι ακολουθίες από bits πάντοτε μπορούν να μετασχηματισθούν σε ακεραίους.

Ωστόσο, υπάρχουν μερικοί απλοί κανόνες που μας παρέχουν τη δυνατότητα να μετατρέψουμε απλούς τύπους σε ακεραίους. Για παράδειγμα, αν τα κλειδιά κατακερματισμού είναι συμβολοσειρές, τότε κάθε χαρακτήρας μετατρέπεται στο ισοδύναμό του κατά ASCII ή Unicode, δηλαδή ένα μικρό ακέραιο. Αθροίζουμε τους ακεραίους και διαιρούμε δια B. Εφόσον το B είναι μικρότερο από το σύνηθες αποτέλεσμα της άθροισης των κωδικών των χαρακτήρων για τον πληθυσμό των συμβολοσειρών, η κατανομή στους κάδους θα είναι σχετικά ομοιόμορφη. Αν η τιμή του B είναι μεγαλύτερη, τότε μπορούμε να επιμερίσουμε τους χαρακτήρες της συμβολοσειράς σε ομάδες μερικών χαρακτήρων η καθεμία. Παραθέτοντας τους κωδικούς των χαρακτήρων κάθε ομάδας δημιουργούμε έναν απλό ακέραιο. Οπότε στη συνέχεια, αθροίζουμε τους ακεραίους που προκύπτουν από όλες τις ομάδες συμβολοσειρών και διαιρούμε δια B όπως πριν. Για παράδειγμα, αν η τιμή του B είναι περί το 1 δις, ή 2^{30}, τότε η ομαδοποίηση των χαρακτήρων σε τετράδες θα δώσει ακεραίους με 32 bits. Το άθροισμα μερικών από αυτούς θα κατανεμηθεί αρκετά αμερόληπτα σε ένα δις κάδων.

Για περισσότερο σύνθετους τύπους δεδομένων, μπορούμε να επεκτείνουμε αναδρομικά την ιδέα μετατροπής των συμβολοσειρών σε ακεραίους.

- Για έναν τύπο εγγραφής, όπου το κάθε πεδίο της είναι διαφορετικού τύπου, αναδρομικά μετατρέπουμε την τιμή κάθε πεδίου σε έναν ακέραιο χρησιμοποιώντας μία αριθμητική κατάλληλη για αυτόν τον τύπο. Αθροίζουμε τους ακεραίους όλων των πεδίων και διαιρούμε το ακέραιο άθροισμα δια B.

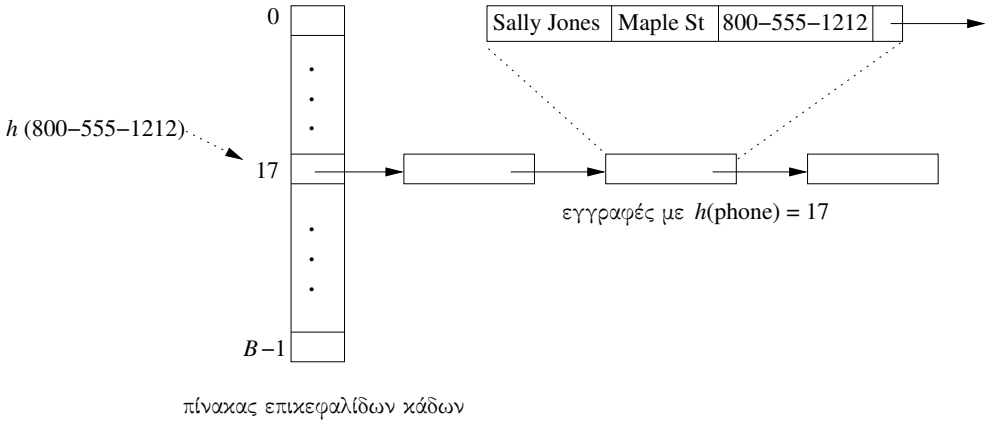
- Για έναν τύπο πίνακα, για έναν τύπο συνόλου ή εν πάσει περιπτώσει για μία ομάδα στοιχείων κάποιου τύπου, μετατρέπουμε τις τιμές των στοιχείων σε ακεραίους, αθροίζουμε τους ακεραίους και διαιρούμε δια B.

1.3.3 Κατάλογοι

Ένας *κατάλογος* είναι μία δομή δεδομένων που καθιστά αποτελεσματική την ανάκτηση αντικειμένων δεδομένης της τιμής ενός ή περισσοτέρων στοιχείων αυτών των αντικειμένων. Η συνηθέστερη κατάσταση είναι όταν τα αντικείμενα είναι εγγραφές, οπότε ο κατάλογος δημιουργείται σε ένα πεδίο αυτής της εγγραφής. Δεδομένης μίας τιμής v για το πεδίο αυτό, ο κατάλογος μας επιτρέπει να ανακτήσουμε όλες τις εγγραφές με τιμή v στο πεδίο αυτό. Για παράδειγμα, θα μπορούσαμε να έχουμε ένα αρχείο με τριπλέτες όνομα-διεύθυνση-τηλέφωνο και έναν κατάλογο για το πεδίο του τηλεφώνου. Δεδομένου ενός τηλεφωνικού αριθμού, ο κατάλογος μας επιτρέπει να βρούμε γρήγορα την εγγραφή ή τις εγγραφές με αυτόν τον τηλεφωνικό αριθμό.

Υπάρχουν πολλοί τρόποι υλοποίησης καταλόγων αλλά δεν θα προσπαθήσουμε να κάνουμε εδώ μία τέτοια επισκόπηση. Οι βιβλιογραφικές σημειώσεις προτείνουν περαιτέρω αναγνώσματα. Ωστόσο, ο πίνακας κατακερματισμού είναι ένας απλός τρόπος δημιουργίας καταλόγων. Το πεδίο ή τα πεδία, όπου βασίζεται ο κατάλογος, σχηματίζουν το κλειδί για τη συνάρτηση κατακερματισμού. Για κάθε εγγραφή εφαρμόζουμε τη συνάρτηση κατακερματισμού στην τιμή του κλειδιού, οπότε η εγγραφή τοποθετείται στον κάδο με διεύθυνση που προσδιορίζεται από τη συνάρτηση κατακερματισμού. Για παράδειγμα, ο κάδος θα μπορούσε να είναι μία λίστα εγγραφών στην κύρια μνήμη ή μία σελίδα στο δίσκο.

Στη συνέχεια, δεδομένης μίας τιμής κλειδιού, μπορούμε να την κατακερματίσουμε, να εντοπίσουμε τον κάδο και να ερευνήσουμε μόνο το περιεχόμενο αυτού του κάδου για να βρούμε εγγραφές με αυτή την τιμή κλειδιού. Αν επιλέξουμε το πλήθος των κάδων B να είναι μία τιμή συγκρίσιμη με το πλήθος των εγγραφών του αρχείου, τότε κάθε κάδος θα περιέχει σχετικά λίγες εγγραφές και ο χρόνος αναζήτησης στον κάδο θα είναι μικρός.



Σχήμα 1.2: *Ένας πίνακας κατακερματισμού χρησιμοποιείται ως κατάλογος. Οι τηλεφωνικοί αριθμοί κατακερματίζονται στους κάδους, ενώ ολόκληρη η εγγραφή τοποθετείται στον κάδο με διεύθυνση που είναι η τιμή κατακερματισμού αυτού του αριθμού.*

Παράδειγμα 1.5: Το Σχήμα 1.2 δείχνει πώς θα έμοιαζε ένας κατάλογος κύριας μνήμης για εγγραφές με όνομα-διεύθυνση-τηλέφωνο. Εδώ, ο κατάλογος δημιουργείται ως προς το πεδίο του τηλεφώνου, ενώ οι κάδοι είναι συνδεδεμένες λίστες. Φαίνεται πως ο αριθμός 800-555-1212 κατακερματίζεται στον κάδο με διεύθυνση 17. Υπάρχει ένας πίνακας με *επικεφαλίδες κάδων*, όπου το i-οστό στοιχείο είναι η κεφαλή μίας συνδεδεμένης λίστας για τον κάδο με διεύθυνση

i. Παρουσιάζεται επίσης η επέκταση ενός στοιχείου της συνδεδεμένης λίστας, το οποίο περιέχει μία εγγραφή με τα πεδία όνομα-διεύθυνση-τηλέφωνο. Στην πραγματικότητα αυτή η εγγραφή είναι που περιέχει τον τηλεφωνικό αριθμό 800-555-1212. Άλλες εγγραφές αυτού του κάδου μπορεί να έχουν αλλά και να μην έχουν αυτόν τον αριθμό. Απλώς γνωρίζουμε ότι οποιοσδήποτε και αν είναι ο τηλεφωνικός αριθμός τους, αυτός κατακερματίζεται στο 17. □

1.3.4 Δευτερεύουσα Αποθήκευση

Όταν χειριζόμαστε δεδομένα μεγάλης κλίμακας, είναι σημαντικό ότι κατανοούμε επαρκώς τις χρονικές διαφορές στην εκτέλεση υπολογισμών όταν τα δεδομένα βρίσκονται αρχικά στο δίσκο, σε αντίθεση με τον απαιτούμενο χρόνο όταν τα δεδομένα βρίσκονται αρχικά στην κύρια μνήμη. Τα φυσικά χαρακτηριστικά των δίσκων είναι ένα άλλο θέμα όπου θα μπορούσαμε να πούμε πολλά αλλά θα πούμε ελάχιστα και θα αφήσουμε τον ενδιαφερόμενο αναγνώστη να ακολουθήσει τις βιβλιογραφικές σημειώσεις.

Οι δίσκοι οργανώνονται σε *σελίδες*, που είναι η ελάχιστη ποσότητα που το λειτουργικό σύστημα χρησιμοποιεί για τη μεταφορά δεδομένων μεταξύ κύριας μνήμης και δίσκου. Για παράδειγμα, το λειτουργικό σύστημα των Windows χρησιμοποιεί σελίδες των 64 KB (δηλαδή, για την ακρίβεια 2^{16}=65.536 bytes). Η προσπέλαση μίας σελίδας του δίσκου (η μετακίνηση της κεφαλής στην άτρακτο της σελίδας και η αναμονή ώστε η σελίδα να περιστραφεί κάτω από την κεφαλή) και η ανάγνωση απαιτούν περίπου 10 msec. Αυτή η καθυστέρηση είναι τουλάχιστον 5 τάξεις μεγέθους (δηλαδή 10^5) μεγαλύτερη από τον απαιτούμενο χρόνο για την ανάγνωση μίας λέξης στην κύρια μνήμη. Συνεπώς, αν αυτό που θέλουμε είναι η προσπέλαση μερικών bytes, τότε το όφελος από το να έχουμε τα δεδομένα στην κύρια μνήμη είναι απαράμιλλο. Στην πραγματικότητα, αν θέλουμε να εκτελέσουμε κάτι απλό σε κάθε byte της σελίδας του δίσκου, π.χ., να χειρισθούμε τη σελίδα σαν έναν κάδο ενός πίνακα κατακερματισμού και να αναζητήσουμε μία συγκεκριμένη τιμή του κλειδιού κατακερματισμού μεταξύ των εγγραφών του κάδου, τότε ο απαιτούμενος χρόνος για τη μετακίνηση της σελίδας από το δίσκο στην κύρια μνήμη θα είναι σαφώς μεγαλύτερος από το χρόνο που απαιτείται για τον υπολογισμό.

Οργανώνοντας τα δεδομένα μας έτσι ώστε τα συσχετιζόμενα δεδομένα να βρίσκονται στον ίδιο *κύλινδρο* (η συλλογή των σελίδων που βρίσκονται σε μία σταθερή ακτίνα από το κέντρο του δίσκου, και συνεπώς είναι προσπελάσιμες χωρίς μετακίνηση της κεφαλής), μπορούμε να αναγνώσουμε όλες τις σελίδες του κυλίνδρου στην κύρια μνήμη σε σημαντικά λιγότερο από 10 msec ανά σελίδα. Μπορούμε να υποθέσουμε ότι ένας δίσκος δεν μπορεί να μεταφέρει δεδομένα

στην κύρια μνήμη με ταχύτητα μεγαλύτερη από 100.000.000 bytes/sec, όπως και να είναι οργανωμένα τα δεδομένα. Αυτό δεν είναι πρόβλημα αν το σύνολο των δεδομένων είναι 1 MB. Όμως ένα σύνολο δεδομένων της τάξης των 100 GB ή 1 TB παρουσιάζει προβλήματα απλώς για την προσπέλαση, πόσο μάλλον για μία χρήσιμη επεξεργασία τους.

1.3.5 Βάση των Φυσικών Λογαρίθμων

Η σταθερά e=2,7182818... έχει μία σειρά χρήσιμων ειδικών ιδιοτήτων. Συγκεκριμένα, το e είναι το όριο του $(1+\frac{1}{x})^x$, όταν το x τείνει στο άπειρο. Οι τιμές αυτής της έκφρασης για $x = 1, 2, 3, 4$ είναι περίπου 2, 2,25, 2,37, 2,44, οπότε γίνεται αποδεκτό ότι το όριο αυτής της σειράς είναι 2,72.

Με αλγεβρικές τεχνικές μπορούμε να προσεγγίσουμε μερικές φαινομενικά σύνθετες εκφράσεις. Ας θεωρήσουμε την έκφραση $(1 + a)^b$, όπου το a λαμβάνει μικρές τιμές. Μπορούμε να ξαναγράψουμε αυτή την έκφραση ως $(1 + a)^{(1/a)(ab)}$. Αντικαθιστούμε $a = 1/x$ και $1/a = x$, οπότε προκύπτει $(1 + \frac{1}{x})^{x(ab)}$ που ισούται με:

$$\left((1 + \frac{1}{x})^x\right)^{ab}$$

Εφόσον το a είναι μικρό, έπεται ότι το x είναι μεγάλο και συνεπώς η έκφραση $(1+\frac{1}{x})^x$ θα πλησιάζει την οριακή τιμή του e. Συνεπώς μπορούμε να προσεγγίσουμε την ποσότητα $(1 + a)^b$ με e^{ab}.

Παρόμοιες ιδιότητες ισχύουν και για αρνητικές τιμές του a. Δηλαδή, για x τείνον στο άπειρο, το όριο του $(1 - \frac{1}{x})^x$ είναι $1/e$. Έπεται ότι η προσέγγιση $(1 + a)^b = e^{ab}$ ισχύει ακόμη και αν το a είναι μικρός αρνητικός αριθμός. Από μία άλλη οπτική γωνία, η έκφραση $(1 - a)^b$ προσεγγίζεται με e^{-ab}, όταν το a είναι μικρό και το b μεγάλο.

Ακολουθούν μερικές ακόμη χρήσιμες προσεγγίσεις με βάση το ανάπτυγμα του Taylor για το e^x, το οποίο είναι:

$$e^x = \sum_{i=0}^{\infty} x^i/i! = 1 + x = x^2/2 + x^3/6 + x^4/24 + \ldots$$

Όταν το x είναι μεγάλο, η σειρά αυτή συγκλίνει αργά, καθώς το $n!$ μεγαλώνει γρηγορότερα από το x^n για κάθε σταθερά x. Ωστόσο, όταν το x είναι μικρό, είτε θετικό είτε αρνητικό, η σειρά συγκλίνει γρήγορα και αρκούν μόνο μερικοί όροι για μία καλή προσέγγιση.

Παράδειγμα 1.6: Εστω ότι $x = 1/2$. Τότε

$$e^{1/2} \;=\; 1 + \frac{1}{2} + \frac{1}{8} + \frac{1}{48} + \frac{1}{384} + \ldots$$

ή προσεγγιστικά $e^{1/2} = 1,64844$.

Εστω ότι $x = -1$. Τότε

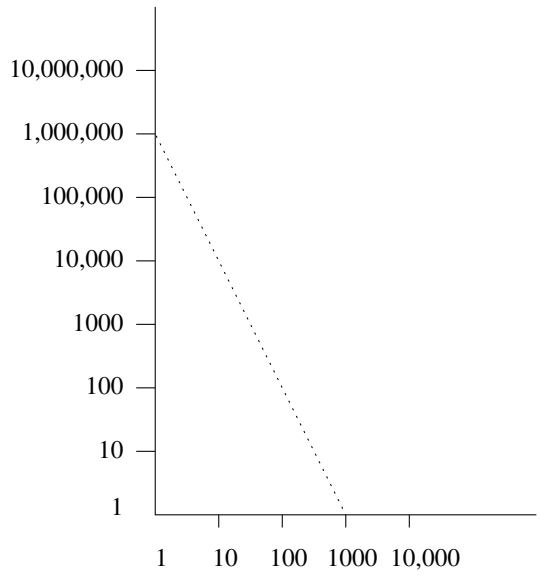
$$e^{-1} \;=\; 1\text{–}1 + \frac{1}{2} - \frac{1}{6} + \frac{1}{24} - \frac{1}{120} + \frac{1}{720} - \frac{1}{5040} + \ldots$$

ή προσεγγιστικά $e^{-1} = 0,36786$. □

1.3.6 Νόμος των Δυνάμεων

Υπάρχουν πολλά φαινόμενα στη φύση που συσχετίζουν 2 μεταβλητές με ένα *Νόμο δυνάμεων*, δηλαδή μία γραμμική σχέση μεταξύ των λογαρίθμων των μεταβλητών. Το Σχήμα 1.3 παρουσιάζει μία τέτοια σχέση. Αν το x είναι ο οριζόντιος και y ο κατακόρυφος άξονας, τότε η σχέση είναι $\log_{10} y = 6 - 2\log_{10} x$.

Παράδειγμα 1.7: Θα μπορούσαμε να εξετάσουμε τις πωλήσεις βιβλίων της Amazon, θεωρώντας ότι η μεταβλητή x παριστά την κατάταξη των βιβλίων ανα-



Σχήμα 1.3: *Ένας Νόμος δύναμης με κλίση -2.*

Η επίδραση του Ματθαίου

Η ύπαρξη Νόμων δυνάμεων με τιμές εκθέτη μεγαλύτερες από 1 εξηγούνται συχνά από την *επίδραση του Ματθαίου*. Στο *Κατά Ματθαίον Ευαγγέλιο* υπάρχει ένας στίχος "οι πλούσιοι γίνονται πλουσιότεροι". Πολλά φαινόμενα παρουσιάζουν αυτή τη συμπεριφορά, καθώς όταν μία ιδιότητα λαμβάνει μία μεγάλη τιμή προκαλεί αυτή την ιδιότητα να επαυξάνεται περαιτέρω. Για παράδειγμα, αν μία ιστοσελίδα έχει πολλούς εισερχόμενους συνδέσμους, τότε είναι πολύ πιθανό πολύ χρήστες να βρουν την ιστοσελίδα αυτή και να επιλέξουν να τη συνδέσουν και από τη δική τους σελίδα. Σαν ένα ακόμη παράδειγμα, αν ένα βιβλίο κάνει καλές πωλήσεις στην Amazon, τότε είναι αναμενόμενο ότι θα διαφημισθεί όταν οι πελάτες επισκέπτονται τον ιστότοπο της Amazon. Κάποιοι από αυτούς τους χρήστες θα επιλέξουν να αγοράσουν επίσης το βιβλίο αυξάνοντας έτσι τις πωλήσεις του.

λόγως των πωλήσεών τους. Τότε y είναι το πλήθος των πωλήσεων του x-οστού πλέον ευπώλητου βιβλίου σε κάποια χρονική περίοδο. Το γράφημα του Σχήματος 1.3 δηλώνει ότι το πλέον ευπώλητο βιβλίο έκανε 1.000.000 πωλήσεις, το 10-το πλέον ευπώλητο έκανε 10.000, ενώ το 100-στό πλέον ευπώλητο έκανε 100 πωλήσεις κ.ο.κ. για όλες τις κατατάξεις μεταξύ και πέραν των αριθμών αυτών. Η συνέπεια ότι πέρα από τη 1000-οστή θέση προκύπτει κλασματικός αριθμός για τις πωλήσεις είναι ακραία. Στην πραγματικότητα θα περιμέναμε ότι η γραμμή εξομαλύνεται για κατατάξεις μεγαλύτερες από 1.000. □

Ο γενικός τύπος του Νόμου δυνάμεων συσχετίζοντας τις μεταβλητές x και y είναι $\log y = b + a \log x$. Αν υψώσουμε τη βάση του λογαρίθμου, ας υποθέσουμε του e (η τιμή αυτή δεν επηρεάζει πραγματικά) με εκθέτες τις δύο πλευρές της εξίσωσης, τότε λαμβάνουμε $y = e^b e^{a \log x} = e^b x^a$. Καθώς e^b είναι απλώς "κάποια σταθερά", την αντικαθιστούμε με σταθερά c. Έτσι, ο Νόμος δυνάμεων μπορεί να γραφεί ως $y = cx^n$ για κάποιες σταθερές a και c.

Παράδειγμα 1.8: Στο Σχήμα 1.3 παρατηρούμε ότι $x = 1, y = 10^6$, ενώ $x = 1.000, y = 1$. Εκτελώντας την πρώτη αντικατάσταση λαμβάνουμε $10^6 = c$. Η δεύτερη αντικατάσταση δίνει $1 = c\ 1000^a$. Εφόσον είναι γνωστό ότι $c = 10^6$, η δεύτερη εξίσωση δίνει $1 = 10^6 1000^a$, οπότε προκύπτει ότι $a = -2$. Συνεπώς ο νόμος που παρουσιάζεται στο Σχήμα 1.3 είναι $y = 10^6 x^{-2}$ ή $y = 10^6/x^2$. □

Στο βιβλίο αυτό θα συναντήσουμε πολλές περιπτώσεις φαινομένων που διέπονται από το Νόμο δυνάμεων. Μερικά παραδείγματα είναι τα εξής:

1. *Βαθμοί Κόμβου σε ένα γράφο του Παγκόσμιου Ιστού*: Διατάσσουμε όλες τις ιστοσελίδες ως προς το πλήθος των εισερχόμενων συνδέσμων τους. Έστω x η θέση μίας ιστοσελίδας σε αυτή την κατάταξη, ενώ y είναι το πλήθος των εισερχόμενων συνδέσμων στη x-οστή ιστοσελίδα. Στην περίπτωση αυτή το y ως συνάρτηση του x μοιάζει πολύ με το Σχήμα 1.3. Ο εκθέτης a είναι λίγο μεγαλύτερος από -2 που παρουσιάζεται στο σχήμα. Έχει βρεθεί ότι προσεγγίζει το 2.1.

2. *Πωλήσεις Προϊόντων*: Διατάσσουμε προϊόντα, ας υποθέσουμε βιβλία της Amazon, ως προς τις πωλήσεις τους το προηγούμενο έτος. Έστω y το πλήθος των πωλήσεων του x-οστού πλέον ευπώλητου βιβλίου. Και πάλι, η συνάρτηση $y(x)$ θα μοιάζει πολύ με το Σχήμα 1.3. Θα συζητήσουμε συνέπειες αυτής της κατανομής των πωλήσεων στην Ενότητα 9.1.2, όπου θα ασχοληθούμε με το θέμα της "μακράς ουράς".

3. *Μεγέθη Ιστοτόπων*: Μετρούμε το πλήθος των σελίδων όλων των ιστοτόπων και τους διατάσσουμε ως προς το πλήθος αυτό. Έστω y το πλήθος των σελίδων του x-οστού ιστοτόπου. Και πάλι, η συνάρτηση $y(x)$ ακολουθεί ένα Νόμο δύναμης.

4. *Ο Νόμος του Zipf*: Αυτός ο Νόμος δύναμης αρχικά αναφερόταν στη συχνότητα των λέξεων σε μία συλλογή εγγράφων. Ας διατάξουμε τις λέξεις ως προς τη συχνότητα και έστω ότι y είναι το πλήθος εμφάνισης της x-οστής λέξης σύμφωνα με την κατάταξη. Τότε λαμβάνουμε ένα Νόμο δύναμης αν και με αρκετά μικρότερη κλίση από αυτή του Σχήματος 1.3. Ο Zipf παρατήρησε ότι ισχύει $y = cx^{-1/2}$. Είναι ενδιαφέρον ότι και άλλα δεδομένα ακολουθούν το συγκεκριμένο Νόμο δύναμης. Για παράδειγμα, αν διατάξουμε τις πολιτείες των ΗΠΑ ως προς τον πληθυσμό τους και παραστήσουμε με y τον πληθυσμό της x-οστής πολιτείας, τότε τα x και y υπακούουν προσεγγιστικά το Νόμο του Zipf.

1.3.7 Ασκήσεις Ενότητας 1.3

Άσκηση 1.3.1: Ας υποθέσουμε ότι ένα αποθετήριο περιέχει 10.000.000 έγγραφα. Ποιά είναι η τιμή (ο πλησιέστερος ακέραιος) του IDF μίας λέξης που εμφανίζεται σε: (α) 40 έγγραφα, και (β) 10.000 έγγραφα;

Άσκηση 1.3.2: Ας υποθέσουμε ότι ένα αποθετήριο περιέχει 10.000.000 έγγραφα, ενώ μία λέξη w εμφανίζεται σε 320 από αυτά. Σε ένα συγκεκριμένο έγγραφο d, το μέγιστο πλήθος εμφανίσεων της λέξης είναι 15. Ποιά είναι προσεγγιστικά η τιμή του TF.IDF της λέξης w αν η λέξη εμφανίζεται: (α) 1 φορά,

και (β) 5 φορές;

Άσκηση 1.3.3 (!)**:** Ας υποθέσουμε ότι επιλέγονται κλειδιά κατακερματισμού από τον πληθυσμό όλων των μη αρνητικών ακεραίων που είναι πολλαπλάσια μίας σταθεράς c, ενώ η συνάρτηση κατακερματισμού είναι $h(x) = x \mod 15$. Για ποιές τιμές της c θα είναι η h μία κατάλληλη συνάρτηση κατακερματισμού, έτσι ώστε ένα μεγάλο σύνολο τυχαίων κλειδιών να κατανέμεται χονδρικά ομοιόμορφα στους κάδους;

Άσκηση 1.3.4: Με βάση το e να δοθούν προσεγγίσεις για το: (α) $1,01^{500}$, (β) $1,05^{1000}$, και (γ) $0,9^{40}$.

Άσκηση 1.3.5: Να χρησιμοποιηθεί το ανάπτυγμα Taylor του e^x για τον υπολογισμό με ακρίβεια 3 δεκαδικών ψηφίων του: (α) $e^{1/10}$, (β) $e^{-1/10}$, και (γ) e^2.

1.4 Περίγραμμα Βιβλίου

Στη συνέχεια ακολουθούν σύντομες περιλήψεις των υπολοίπων κεφαλαίων του βιβλίου.

Το Κεφάλαιο 2 αναφέρεται στην Εξόρυξη Δεδομένων αυτή καθ' εαυτή αλλά μας εισάγει στη μεθοδολογία Map-Reduce για την εκμετάλλευση του παραλληλισμού σε υπολογιστικά νέφη (ράφια-racks διασυνδεδεμένων επεξεργαστών). Υπάρχουν λόγοι να πιστεύουμε ότι η υπολογιστική νέφους, και ειδικότερα η μεθοδολογία Map-Reduce, θα καταστούν ο φυσιολογικός τρόπος υπολογισμού κατά την ανάλυση μεγάλων ποσοτήτων δεδομένων. Ένα σημαντικό θέμα σε επόμενα κεφάλαια είναι η εκμετάλλευση της μεθοδολογίας Map-Reduce για την υλοποίηση των διαφόρων αλγορίθμων.

Το Κεφάλαιο 3 αναφέρεται στην εύρεση παρόμοιων ειδών. Σημείο αφετηρίας είναι ότι τα είδη μπορούν να αναπαρασταθούν ως σύνολα αντικειμένων, οπότε παρόμοια σύνολα είναι εκείνα που έχουν κοινό μεγάλο κλάσμα των αντικειμένων τους. Εξηγούνται οι βασικές τεχνικές του ελάχιστου κατακερματισμού (minhashing) και του ευαίσθητου σε τοπικότητα κατακερματισμού (locality-sensitive hashing). Οι τεχνικές αυτές έχουν πληθώρα εφαρμογών και συχνά δίνουν απροσδόκητα αποτελεσματικές λύσεις σε προβλήματα που εμφανίζονται δυσεπίλυτα για μαζικά σύνολα δεδομένων.

Στο Κεφάλαιο 4 θεωρούμε δεδομένα υπό τη μορφή ροής . Η διαφορά μεταξύ μίας ροής δεδομένων και μίας βάσης δεδομένων είναι ότι τα δεδομένα της ροής χάνονται αν δεν κάνουμε αμέσως κάτι για αυτά. Χαρακτηριστικά παραδείγματα ροών είναι οι ροές των ερωτημάτων σε μία μηχανή αναζήτησης ή τα κλικ σε ένα

δημοφιλή ιστότοπο. Στο κεφάλαιο αυτό θα εξετάσουμε διάφορες εκπληκτικές εφαρμογές του κατακερματισμού που καθιστούν εφικτή τη διαχείριση των ροών.

Το Κεφάλαιο 5 είναι αφιερωμένο σε μία μοναδική εφαρμογή, τον υπολογισμό του PageRank. Αυτός ο υπολογισμός είναι η ιδέα που έκανε τη Google να ξεχωρίσει από τις άλλες μηχανές αναζήτησης και είναι ακόμη το ουσιαστικότερο κομμάτι των μηχανών αυτών για να γνωρίζουμε ποιές σελίδες είναι πιθανότερο να θέλουν να δουν οι χρήστες. Επεκτάσεις του PageRank είναι επίσης ουσιαστικές στη μάχη εναντίον των ανεπιθύμητων ηλεκτρονικών μηνυμάτων spam (οι οποίες κατ'ευφημισμό ονομάζονται "βελτιστοποίηση των μηχανών αναζήτησης"), οπότε θα εξετάσουμε τις τελευταίες επεκτάσεις της ιδέας αυτής για την καταπολέμηση του spam.

Στη συνέχεια, το Κεφάλαιο 6 εισάγει το μοντέλο δεδομένων, το οποίο ονομάζεται "καλάθι αγορών" και τα σχετικά προβλήματα των κανόνων συσχέτισης και της εύρεσης συχνών συνόλων ειδών. Στο μοντέλο του καλαθιού αγορών, τα δεδομένα αποτελούνται από μία μεγάλη συλλογή καλαθιών, όπου το καθένα περιέχει ένα μικρό σύνολο ειδών. Δίνουμε μία σειρά αλγορίθμων για την εύρεση όλων των συχνών ζευγών ειδών, δηλαδή ζευγών ειδών που εμφανίζονται μαζί σε πολλά καλάθια. Μία άλλη σειρά αποτελεσματικών αλγορίθμων χρησιμοποιείται για την εύρεση των περισσότερων από τα συχνά σύνολα ειδών, τα οποία είναι μεγαλύτερα του ζεύγους.

Το Κεφάλαιο 7 εξετάζει το πρόβλημα της Ομαδοποίησης. Υποθέτουμε ένα σύνολο ειδών και ένα μέτρο απόστασης, το οποίο ορίζει πόσο κοντά ή μακριά είναι δύο είδη μεταξύ τους. Σκοπός είναι η εξέταση μεγάλων όγκων δεδομένων και ο διαμερισμός τους σε ομάδες (clusters), όπου κάθε ομάδα αποτελείται από είδη που είναι πολύ κοντά μεταξύ τους αλλά ταυτοχρόνως είναι μακριά από είδη των άλλων ομάδων.

Το Κεφάλαιο 8 είναι αφιερωμένο στην άμεση διαφήμιση και τα υπολογιστικά προβλήματα που αυτή γεννά. Εισάγουμε την έννοια ενός άμεσου αλγορίθμου, όπου πρέπει να δοθεί αμέσως μία καλή απάντηση αντί να περιμένουμε μέχρι να μας δοθεί ολόκληρο το σύνολο δεδομένων. Η ιδέα του ανταγωνιστικού λόγου είναι μία ακόμη σημαντική έννοια που καλύπτεται στο κεφάλαιο αυτό. Είναι η αναλογία της εγγυημένης επίδοσης ενός άμεσου αλγορίθμου σε σύγκριση με την επίδοση ενός βέλτιστου αλγορίθμου στον οποίο επιτρέπεται να γνωρίζει όλα τα δεδομένα πριν καταλήξει σε απόφαση. Αυτές οι ιδέες είναι εύκολο να δώσουν καλούς αλγορίθμους που ταιριάζουν τις προσφορές των διαφημιστών (για το δικαίωμα να επιδείξουν τη διαφήμισή τους σε απάντηση μίας ερώτησης) έναντι των ερωτημάτων αναζήτησης που φθάνουν σε μία μηχανή αναζήτησης.

Τέλος, το Κεφάλαιο 9 είναι αφιερωμένο στα συστήματα συστάσεων. Πολλές

εφαρμογές του Παγκόσμιου Ιστού εμπεριέχουν τη συμβουλή προς τους χρήστες σχετικά με το τι θα τους άρεσε. Η πρόκληση του Netflix είναι ένα παράδειγμα, όπου επιθυμούμε να προβλέψουμε τί ταινίες θα άρεσαν σε ένα χρήστη ή το πρόβλημα της Amazon να πλασάρει ένα προϊόν σε ένα πελάτη με βάση πληροφορίες σχετικά με το ενδιαφέρον του για αγορές. Υπάρχουν δύο βασικές προσεγγίσεις για τις συστάσεις. Μπορούμε να χαρακτηρίσουμε τα είδη με γνωρίσματα, όπως οι πρωταγωνιστές ενός φιλμ, και να συστήσουμε είδη με γνωρίσματα ίδια με αυτά που είναι γνωστό ότι αρέσουν στο χρήστη. Ή, μπορούμε να δούμε επιλογές άλλων χρηστών με προτιμήσεις παρόμοιες με αυτές του υπόψη χρήστη (μία τεχνική γνωστή ως συνεργατική διήθηση).

1.5 Περίληψη Κεφαλαίου 1

- *Εξόρυξη Δεδομένων*: Αυτός ο όρος αναφέρεται στη διαδικασία εξαγωγής χρήσιμων μοντέλων δεδομένων. Μερικές φορές, ένα μοντέλο μπορεί να είναι μία περίληψη των δεδομένων ή μπορεί να είναι το σύνολο των πιο ακραίων γνωρισμάτων των δεδομένων.

- *Αρχή Bonferroni:* Αν είμαστε πρόθυμοι να θεωρήσουμε σαν σημαντικό γνώρισμα των δεδομένων κάτι που αναμένεται να παρουσιασθεί πολλές φορές σε τυχαία δεδομένα, τότε δεν μπορούμε να βασιζόμαστε στο ότι τέτοια γνωρίσματα είναι σημαντικά. Αυτή η παρατήρηση περιορίζει την ικανότητά μας να εξορύξουμε δεδομένα για γνωρίσματα που δεν είναι επαρκώς σπάνια στην πράξη.

- *TF.IDF*: Το μέτρο TF.IDF μας επιτρέπει να αναγνωρίσουμε λέξεις σε μία συλλογή εγγράφων, οι οποίες είναι χρήσιμες για τον προσδιορισμό του θέματος του κάθε εγγράφου. Μία λέξη έχει υψηλή τιμή TF.IDF σε ένα έγγραφο αν εμφανίζεται σε σχετικά λίγα έγγραφα αλλά εμφανίζεται και στο συγκεκριμένο, ενώ όταν εμφανίζεται σε ένα έγγραφο τείνει να εμφανίζεται πολλές φορές.

- *Συναρτήσεις Κατακερματισμού*: Μία συνάρτηση κατακερματισμού απεικονίζει τα κλειδιά κάποιων τύπων δεδομένων σε ακέραιες διευθύνσεις κάδων. Μία καλή συνάρτηση κατανέμει ομοιόμορφα όλες τις εν δυνάμει τιμές κλειδιών στους κάδους. Οποιοσδήποτε τύπος δεδομένων μπορεί να είναι το πεδίο μίας συνάρτησης κατακερματισμού.

- *Κατάλογοι*: Ένας κατάλογος είναι μία δομή δεδομένων που μας επιτρέπει να αποθηκεύσουμε και να ανακτήσουμε εγγραφές αποτελεσματικά, δεδομένης

της τιμής ενός ή περισσοτέρων πεδίων της εγγραφής. Ο κατακερματισμός είναι ένας τρόπος δημιουργίας καταλόγων.

- *Αποθήκευση στο Δίσκο*: Όταν τα δεδομένα πρέπει να αποθηκευθούν στο δίσκο (δευτερεύουσα μνήμη), απαιτείται πολύ περισσότερος χρόνος για την προσπέλαση των επιθυμητών δεδομένων σε σύγκριση με την περίπτωση που τα ίδια δεδομένα είναι αποθηκευμένα στην κύρια μνήμη. Όταν τα δεδομένα είναι ογκώδη, είναι απαραίτητο να γίνει προσπάθεια ώστε οι αλγόριθμοι να διατηρήσουν τα απαραίτητα δεδομένα στην κύρια μνήμη.

- *Νόμοι Δυνάμεων*: Πολλά φαινόμενα υπακούουν σε κανόνες που μπορούν να εκφρασθούν ως $y = cx^a$ για κάποιον εκθέτη a, συχνά περί το -2. Τέτοια φαινόμενα συμπεριλαμβάνουν τις πωλήσεις του x-οστού πλέον ευπώλητου βιβλίου, ή το πλήθος των εισερχόμενων συνδέσμων στην x-οστή πλέον δημοφιλή ιστοσελίδα.

1.6 Αναφορές Κεφαλαίου 1

Το βιβλίο [7] αποτελεί μία σαφή εισαγωγή στα βασικά της Εξόρυξης Δεδομένων. Το βιβλίο [2] καλύπτει την Εξόρυξη Δεδομένων κυρίως από την οπτική γωνία της Μηχανικής Μάθησης και της στατιστικής. Για την κατασκευή συναρτήσεων κατακερματισμού και πινάκων κατακερματισμού, βλέπε το βιβλίο [4]. Λεπτομέρειες για το μέτρο TF.IDF και άλλα θέματα σχετικά με την επεξεργασία εγγράφων μπορούν να βρεθούν στο βιβλίο [5]. Στο βιβλίο [3] υπάρχουν περισσότερα σχετικά με τη διαχείριση καταλόγων, πινάκων κατακερματισμού και δεδομένων στο δίσκο. Οι Νόμοι δυνάμεων που σχετίζονται με τον Παγκόσμιο Ιστό διερευνήθηκαν στην εργασία [1]. Η επίδραση του Ματθαίου παρατηρήθηκε για πρώτη φορά στην εργασία [6].

1. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Weiner, "Graph structure in the web", *Computer Networks*, Vol.33, No.1-6, pp.309-320, 2000.

2. M.M. Gaber, *Scientific Data Mining and Knowledge Discovery – Principles and Foundations*, Springer, New York, 2010.

3. H. Garcia-Molina, J.D. Ullman, and J. Widom, *Database Systems: The Complete Book*, 2nd Edition, Prentice-Hall, Upper Saddle River, NJ, 2009.

4. D.E. Knuth, *The Art of Computer Programming, Vol. 3 (Sorting and Searching)*, 2nd Edition, Addison-Wesley, Upper Saddle River, NJ, 1998.

5. C.P. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

6. R.K. Merton, "The Matthew effect in science", *Science*, Vol.159, No.3810, pp.56-63, Jan.5, 1968.

7. P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Addison-Wesley, Upper Saddle River, NJ, 2005.