ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΚΩΔΙΚΟΠΟΙΗΣΗ ΠΗΓΗΣ ΣΕ ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΟ ΠΕΡΙΒΑΛΛΟΝ

ΦΟΙΤΗΤΗΣ: ΝΤΕΝΕΖΟΣ ΤΑΞΙΑΡΧΗΣ ΜΙΧΑΗΛ (ΑΜ 2698)

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΠΑΡΑΣΚΕΥΑΣ ΜΙΧΑΛΗΣ

ΗΜΕΡΟΜΗΝΙΑ: ΙΟΥΝΙΟΣ 2023

Πίνακας Περιεχομένων

Πίνακας Εικόνων

Πίνακας Διαγραμμάτων

Πίνακας Πινάκων

Ευχαριστίες

Στο σημείο αυτό, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες αρχικά προς τον κύριο Παρασκευά Μιχάλη, καθηγητή του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Πελοποννήσου, για την αμέριστη βοήθεια, κατανόηση και υπομονή στην διάρκεια τόσο της πτυχιακής μου εργασίας όσο και του κύκλου σπουδών μου. Επιπλέον, θα ήθελα να ευχαριστήσω ιδιαίτερα το ακαδημαϊκό προσωπικό και όλους τους καθηγητές μου που συνέβαλλαν πρόθυμα στην καθοδήγηση και την παροχή γνώσεων για τον πολύ ενδιαφέρον κλάδο της πληροφορικής. Τέλος, θα ήθελα να πω ένα τεράστιο ευχαριστώ προς την οικογένεια μου και την Ευαγγελία για τη συμπαράσταση και την παρότρυνση τους να ακολουθήσω τις ακαδημαϊκές μου σπουδές και να πετύχω τον πρώτο μεγάλο στόχο ζωής μου.

Πρόλογος

Η Θεωρία Πληροφορίας θεωρείται υψίστης σημασίας κλάδος που εκτείνεται σε ένα μεγάλο πλήθος επιστημονικών πεδίων όπως είναι η πληροφορική, τα οικονομικά, η στατιστική, η φυσική, η μηχανική και διάφορα ακόμη. Η πρώτη επιστημονική προσέγγιση της πληροφορίας ως ποσοτικοποιημένη έννοια πραγματοποιήθηκε από τον Hartley το 1928. Με την πάροδο των ετών και χάρη στο μεγαλειώδης έργο του Claude E. Shannon πάνω στην επεξεργασία σήματος, ξεκίνησε η μαθηματική ολοκλήρωση της θεωρίας.

Χάρη στην εξέλιξη της Θεωρία της Πληροφορίας, η οποία είναι βασισμένη στην στατιστική, την θεωρία πιθανοτήτων και την άλγεβρα, κρίθηκε ευκολότερο να απαντηθούν ουσιώδης ερωτήματα σχετικά με την βέλτιστη συμπίεση δεδομένων, την περιγραφή των διαύλων επικοινωνίας, την κωδικοποίηση μηνυμάτων πληροφορίας, το ρυθμό μετάδοσης των πληροφοριών, την κρυπτογράφηση και πολλά ακόμη. Ως συνέπεια των παραπάνω ερωτημάτων και για πρακτικούς λόγους διαμορφώθηκε η Θεωρία Κωδικοποίησης, η οποία αποτελεί εκείνο το επιστημονικό θεώρημα που μελετά τις μεθόδους αποτελεσματικής μεταφοράς της πληροφορίας από την πηγή στον προορισμό της. Στην παρούσα εργασία, θα γίνει μια περιγραφική ανάλυση της Θεωρίας Πληροφορίας περιγράφοντας τις ρίζες δημιουργίας της, τους επιστήμονες που συνεισφέραν στη διαμόρφωση της, καθώς και όλες τις βασικές πτυχές που την περιγράφουν σήμερα.

Στο πρώτο κεφάλαιο, το οποίο αποτελεί εισαγωγικό κεφάλαιο, αναφέρονται συνοπτικά η ιστορική αναδρομή της έννοιας της πληροφορίας καθώς και περιγραφικές έννοιες, οι περισσότερες από τις οποίες θα εξελιχθούν στην συνέχεια. Το δεύτερο κεφάλαιο αναφέρεται στις μαθηματικές βάσεις, όπως την θεωρία πιθανοτήτων, το θεώρημα του Bayes, το μέτρο πληροφορίας του Hartley, τη μέση ποσότητα πληροφορίας (ή εντροπία) του Shannon και τις έννοιες της συνδυασμένης, της υπό συνθήκης και της αμοιβαίας πληροφορίας. Τέλος, στο τρίτο και τελευταίο κεφάλαιο θα γίνει μια επικεντρωμένη ανάλυση των Διακριτών Πηγών αναλύοντας τις δύο διαφορετικές πτυχές που υφίστανται, τις Διακριτές Πηγές χωρίς μνήμη και με μνήμη.

ΚΕΦΑΛΑΙΟ 1

Η ΠΛΗΡΟΦΟΡΙΑ ΣΕ ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΟ ΠΕΡΙΒΑΛΛΟΝ

1.1 Εισαγωγή

Η έννοια της πληροφορίας *(Information)* συνδέεται έντονα με τη δυνατότητα ανακάλυψης και μετάδοσης γνώσης. Συχνά, η πληροφορία ταυτίζεται με την έννοια των δεδομένων *(Data)* καθώς αποτελούν το πρώτο στάδιο της γέννησης της. Τα δεδομένα δεν είναι πληροφορία όταν δεν έχουν λάβει από τη νόηση συγκεκριμένη σημασία, ενώ αντίθετα η πληροφορία είναι δεδομένα με ουσιαστική σημασία και περιεχόμενο. Σε γενικές γραμμές, η πληροφορία είναι η γνώση που όταν φτάσει από τον πομπό στο δεκτή επηρεάζει με μη τυχαίο τρόπο την εξέλιξη οποιουδήποτε συστήματος.

Η έννοια της πληροφορίας στη κοινή ορολογία δεν διαφέρει αισθητά σε ένα προγραμματιστικό περιβάλλον. Στη στατιστική περιγραφή του όρου προερχόμενη από το εφαρμοσμένο επιστημονικό πεδίο της ηλεκτρολογίας και της μετάδοσης δεδομένων, σχετίζεται με τη μαθηματική θεωρία, γνωστή ως Θεωρίας της Πληροφορίας *(Information Theory)*. Η Θεωρία Πληροφορίας είναι κλάδος της Θεωρίας Πιθανοτήτων *(Probability Theory)* και σκοπό έχει την ακριβή ποσοτικοποίηση της πληροφορίας. Ο κλάδος αυτός αναπτύχθηκε αρχικά από τον Claude E. Shannon με την εργασία του, Μαθηματική Θεωρία Επικοινωνίας (1948) πάνω στην επεξεργασία σήματος. Συγκεκριμένα, αναφέρθηκε στη διαδοχική επιλογή συμβόλων ή λέξεων από ένα προκαθορισμένο αλφάβητο προκειμένου να αποσταλεί ένα μήνυμα ή νόημα. Ως αποτέλεσμα του έργου του, διαπιστώθηκε η δυνατότητα μετάδοσης της πληροφορίας βασιζόμενη στη θεωρία πιθανοτήτων και σε ουσιαστικά επιμέρους στατιστικά εργαλεία.

Σε αυτό το κεφάλαιο θα γίνει μια σύντομη περίληψη της εξέλιξης της θεωρίας της πληροφορίας καθώς και των επιστημόνων που συνέβαλαν στην δημιουργία της με τα έργα τους. Στην συνέχεια, θα γίνει μια ταξινόμηση των βασικότερων κατηγοριών που θα μπορούσε να διακριθεί η έννοια της πληροφορίας. Τέλος, θα αναφερθεί εν συντομία η διαδικασία κωδικοποίησης της πληροφορίας από την πηγή στον προορισμό μέσα σε ένα προγραμματιστικό περιβάλλον.

1.2 Ιστορική Αναδρομή της Θεωρίας της Πληροφορίας

Αν και η πρώτη επιστημονική προσέγγιση της πληροφορίας ως ποσοτικοποιημένη έννοια πραγματοποιήθηκε από τον Hartley το 1928, η διαπίστωση της φύσης της λέξης ως συμπέρασμα μέσω δεδομένων υπήρχε από αρχαιοτάτων χρονών. Η προσπάθεια του ανθρώπου των σπηλαίων να μεταφέρει την πληροφορία εντοπισμού τροφής σε έναν συνάνθρωπο του μέσω της γλώσσας του σώματος αποτελεί χαρακτηριστικό παράδειγμα.

Με το πέρασμα των ετών και χάρη στο έργο του Hartley και στη συνέχεια του Shannon, καθορίστηκε τόσο η έννοια της πληροφορίας όσο και η ποσοτικοποίηση της. Πρώτος ο Hartley το 1928 διαπίστωσε την αδυναμία ποσοτικοποίησης της πληροφορίας και αρχικά επιχείρησε να την ορίσει ενώ στη συνέχεια, να την τεκμηριώσει μαθηματικά. Επομένως, όρισε την ποσότητα της πληροφορίας ως το δεκαδικό λογάριθμο του πλήθους των διαφορετικών λέξεων που μπορούν να σχηματιστούν, αποτελούμενες από ένα δεδομένο πλήθος συμβόλων. Στην περίπτωση μηνυμάτων μήκους k συμβόλων από ένα αλφάβητο με N σύμβολα η ποσότητα πληροφορίας είναι ίση με:

$$H(N^k) = log(N^k) = k\, log\, N$$

Καθοριστικό ρόλο στην ανάπτυξη της Θεωρίας της Πληροφορίας είχε ο Shannon, ο οποίος το 1948 έθεσε τις βάσεις του στο επιστημονικό του άρθρο με τίτλο *«Μαθηματική Θεωρία Επικοινωνίας» (A Mathematical Theory of Communication)*. Στόχος της σημαντικής αυτής θεωρίας είναι η θεμελίωση εννοιών και θεωρημάτων που επιτρέπουν τη μαθηματική περιγραφή της διαδικασίας της επικοινωνίας. Με άλλα λόγια, η μετάδοση πληροφοριών μπορεί να αναλυθεί με μαθηματική αυστηρότητα και ακρίβεια ενώ μελλοντικά είναι δυνατόν να σχεδιαστούν καλύτερα συστήματα επικοινωνίας. Η σημαντικότερη έννοια που ασχολήθηκε ο Shannon αφορούσε τη μέση ποσότητα πληροφορίας, γνωστή και ως εντροπία. Η εν λόγο έννοια, αποτελεί θεμελιώδης βάση στη θεωρία της πληροφορίας και θα αναλυθεί εκτενέστερα στο επόμενο κεφάλαιο.

Με την Θεωρία της Πληροφορίας, η οποία είναι βασισμένη στην στατιστική, την θεωρία πιθανοτήτων και την άλγεβρα, μπορούν να απαντηθούν ερωτήματα σχετικά με την βέλτιστη συμπίεση δεδομένων, την περιγραφή των διαύλων επικοινωνίας, την κωδικοποίηση μηνυμάτων πληροφορίας, το ρυθμό μετάδοσης των πληροφοριών σε περιβάλλον θορύβου, την κρυπτογράφηση και πολλά ακόμη. Εν ολίγης, η θεωρία της

πληροφορίας ασχολείται με τα μέτρα και τις εφαρμογές της έννοιας της πληροφορίας και απαντά στις εξής δύο πολύ βασικές ερωτήσεις:

- ✓ Ποια είναι η μεγαλύτερη δυνατή συμπίεση δεδομένων και
- ✓ ποιος είναι ο μέγιστος δυνατός ρυθμός μετάδοσης σε ένα επικοινωνιακό κανάλι.

Οι δύο ερωτήσεις προσπαθούν να εντοπίσουν τόσο τη μέγιστη δυνατή αφαίρεση πλεονασμού περιεχομένου, με σκοπό τη συμπιεσμένη αναπαράσταση των μηνυμάτων, όσο και τον ρυθμό μέσα στον οποίο είναι εφικτό αυτό το γεγονός. Η πρώτη ερώτηση μπορεί να απαντηθεί μέσω της εντροπίας ή μέσης ποσότητας πληροφορίας ενώ η δεύτερη ερώτηση μέσω της χωρητικότητας του καναλιού.

Η εξέλιξη της Θεωρίας της Πληροφορίας καθορίστηκε μέσω σπουδαίας διαχρονικής προσπάθειας διαφόρων επιστημόνων με επικρατέστερους τον Ralph V. R. Hartley και τον Claude E. Shannon. Ο Hartley πρώτος διαπίστωσε την ανάγκη ποσοτικοποίησης της έννοιας της πληροφορίας και ο Shannon στην συνέχεια, οραματίστηκε και εξέλιξε την εν λόγο έννοια με τη εισαγωγή της Θεωρίας Πιθανοτήτων θέτοντας τις βάσεις της σύγχρονης Θεωρίας Πληροφορίας.

1.3 Έννοιες της Πληροφορίας

Η επιστημονική προσέγγιση που αφορά την έννοια της πληροφορίας ονομάστηκε «Θεωρία της Πληροφορίας» και αποτελεί ένα τεκμηριωμένα μαθηματικό κομμάτι της πληροφοριακής επιστήμης. Συγκεκριμένα, πρόκειται για εκείνο το πεδίο που ασχολείται με τον ορισμό της έννοιας, τον προσδιορισμό των μέτρων και των εφαρμογών που την αποτελούν. Θα ήταν εφικτό να ταξινομηθούν οι επικρατέστερες έννοιες της πληροφορίας σε τρεις διαφορετικές κατηγορίες ως εξής:

- **Συντακτική Πληροφορία**

Ένα μήνυμα έχει συγκεκριμένο νόημα *(Syntactic Information)* και διαμορφώνεται μέσω κάποιου συστήματος με καθορισμένες φυσικές ή εννοιολογικές οντότητες. Οι σημασιολογικές αυτές οντότητες αποτελούν τελείως διαφορετικό κομμάτι από τη μηχανική λειτουργία. Με άλλα λόγια, το αποτέλεσμα του μηνύματος είναι επιλεγμένο από ένα σύνολο πιθανών μηνυμάτων, λέξεων ή συμβόλων. Το σύστημα πρέπει να είναι σχεδιασμένο για να λειτουργεί για κάθε πιθανή επιλογή, όχι μόνο για αυτήν που θα επιλεγεί κάθε φορά, καθώς αυτό είναι άγνωστο τη στιγμή του σχεδιασμού. Η

συντακτική πληροφορία σχετίζεται με τα σύμβολα και τις σχέσεις μεταξύ αυτών, τα οποία επιλέγονται τυχαία, αποτελούν και εν τέλει διαμορφώνουν ένα μήνυμα με συγκεκριμένο νόημα.

- **Σημασιολογική Πληροφορία**

Η σημασιολογική πληροφορία *(Semantic Information)* σχετίζεται με τη σημασία ή το νόημα που μεταδίδει ένα μήνυμα. Εν ολίγης, δύο προτάσεις είναι πιθανό να διαφέρουν ως προς το νόημα που μεταδίδουν ασχέτως μεγέθους λέξεων ή συμβόλων.

- **Πραγματική Πληροφορία**

Η πραγματική πληροφορία *(Real Information)* ταυτίζεται με τη χρήση και τη δυνατή επίπτωση των μηνυμάτων που έχουν αποσταλεί. Για παράδειγμα, διαπιστώνεται έντονη διαφορά στη χρήση ενός μηνύματος που απευθύνεται σε ένα συγκεκριμένο πλαίσιο όταν αυτός που το λαμβάνει συνδέεται άμεσα με εκείνο ή αντίθετα.

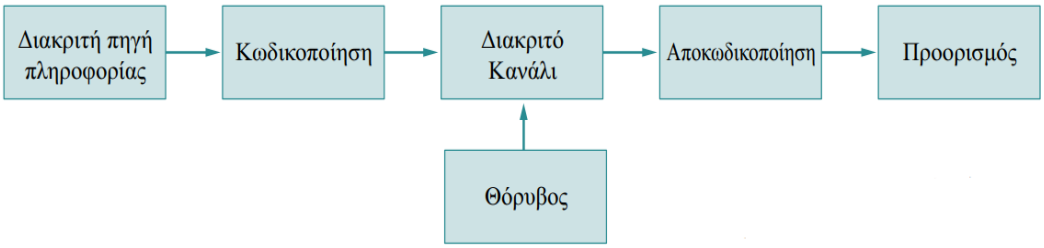
Στο σύνολο των θεμάτων που απασχολούν τη Θεωρία Πληροφορίας εντάσσονται η ποσότητα συντακτικής πληροφορίας (ή εντροπία) και οι μονάδες μέτρησης αυτής, καθώς και η ροή πληροφορίας σε κανάλια. Επιπλέον, εμπεριέχει τα θεμελιώδη όρια της ποσότητας πληροφορίας που μπορούν να μεταδοθούν, δηλαδή η χωρητικότητα καναλιών, που αποτελεί το μέγγιστο δυνατό ρυθμό μετάδοσης. Στο πλαίσιο αυτής της εργασίας θα αναφερθούν μόνο τα επικρατέστερα σημεία της θεωρίας που αφορούν τη συντακτική πληροφορία, δηλαδή τη πληροφορία που εξαρτάται από την πιθανότητα εμφάνισης των μηνυμάτων και όχι από τη σημασία τους.

1.4 Θεωρία Κωδικοποίησης Πληροφορίας στη Πηγή

Η κωδικοποίηση χαρακτήρων *(Coding Theory)* είναι η διαδικασία αντιπροσώπευσης μεμονωμένων χαρακτήρων χρησιμοποιώντας ένα αντίστοιχο σύστημα κωδικοποίησης που αποτελείται από άλλα σύμβολα και τύπους δεδομένων, ενώ χρησιμοποιείται για ποικίλους σκοπούς. Βασικός στόχος η διευκόλυνση αποθήκευσης, διαχείρισης κειμένου σε υπολογιστικά συστήματα καθώς και μεταφοράς κειμένου μέσω τηλεπικοινωνιακών δικτύων.

Η Θεωρία Κωδικοποίησης αποτελεί εκείνο το επιστημονικό θεώρημα που μελετά τις μεθόδους αποτελεσματικής μεταφοράς της πληροφορίας από την πηγή στον προορισμό

της. Το φυσικό μέσο, δια του οποίου μεταδίδεται η πληροφορία, καλείται κοινώς κανάλι επικοινωνίας με σκοπό να φτάσει ένα ορθός νοηματικά μήνυμα στον προορισμό. Στην Εικόνα 1, γίνεται οπτικά κατανοητή η διαδικασία μεταφοράς μίας πληροφορίας.



Πηγή: Θεωρία Πληροφορίας και Κωδικοποίησης, Βασίλειος Ζορκάδης (2002)

Εικόνα 1 Διακριτό Κανάλι Επικοινωνίας

Η πηγή παράγει πληροφορία σε μορφή που δεν είναι κατάλληλη για την άμεση μετάδοση της μέσω του καναλιού. Για το λόγο αυτό, η πληροφορία υποβάλλεται σε ειδική επεξεργασία, την επονομαζόμενη «κωδικοποίηση του καναλιού» ενώ προηγουμένως έχει προηγηθεί «κωδικοποίηση πηγής» με σκοπό την απομάκρυνση του πλεονασμού από την έξοδο της πηγής. Η πηγή όπου παρέχει τη πληροφορία παράγει ακολουθίες συμβόλων (ή γραμμάτων), όπου το σύνολο των συμβόλων ονομάζεται αλφάβητο πηγής ενώ μια ομάδα διαδοχικών συμβόλων καλείται μήνυμα ή λέξη. Τα σύμβολα δημιουργούνται στη πηγή σε διακριτές χρονικές στιγμές. Συμπερασματικά, οι εν λόγο πηγές καλούνται διακριτές πηγές πληροφορίας και διακρίνονται σε δύο κατηγορίες ανάλυσης, εκείνες χωρίς μνήμη και εκείνες με μνήμη.

Η Θεωρία Κωδικοποίησης, όπως αναφέρθηκε, μπορεί να εκτελεστεί είτε στην πηγή προκειμένου να αφαιρεθούν τυχόν πλεονασμοί κατά την έξοδο αλλά και στο κανάλι για να γίνει η απαιτούμενη ειδική επεξεργασία. Δεδομένου του γεγονότος πως η διαδικασία επικοινωνίας από τη πηγή στο προορισμό αποτελείται από αρκετά μεγάλο φάσμα εννοιών και επιμέρους διαδικασιών, σε αυτή την εργασία θα αναλυθεί μόνο το πρώτο στάδιο που αφορά τη κωδικοποίηση διακριτών πηγών χωρίς και με μνήμη, καθώς και τους σχετιζόμενους αλγορίθμους κωδικοποίησης.

1.5 Ανακεφαλαίωση

Στο κεφάλαιο αυτό περιγράφηκε η ιστορική εξέλιξη της θεωρίας της πληροφορίας αναφέροντας τους συμβαλλόμενους επιστήμονες που αποτέλεσαν σπουδαίο ρόλο στην σημερινή της διαμόρφωση. Παράλληλα, αναλύθηκε η έννοια της πληροφορίας σε τρείς διαφορετικές κατηγορίες, κάθε μια από τις οποίες έχει διαφορετικό νόημα και στόχο. Τέλος, περιγράφηκε περιληπτικά η διαδικασία μετάδοσης της πληροφορίας, ο ορισμός των διακριτών πηγών επικοινωνίας μέσα στα οποία διαμορφώνεται ένα μήνυμα καθώς και η έννοια της κωδικοποίησης στο στάδιο της πηγής.

ΚΕΦΑΛΑΙΟ 2

ΒΑΣΙΚΕΣ ΣΤΑΤΙΣΤΙΚΕΣ ΕΝΝΟΙΕΣ ΠΛΗΡΟΦΟΡΙΑΣ

2.1 Εισαγωγή

Για πολλά χρόνια η πιθανότητα ενός ευνοϊκού ή μη γεγονότος ονομαζόταν τύχη. Το 1225-1274 μ.Χ., ο Thomas Aquinas, ένας Ιταλός φιλόσοφος και θεολόγος θεώρησε ότι ορισμένα γεγονότα ονομάζονται τυχαία διότι δεν γίνεται να συλλεχθούν όλα τα εμπλεκόμενα δεδομένα για να ερμηνευτεί μια κατάσταση. Το 17ο αιώνα, χάρη στο ενδιαφέρον του ανθρώπου για τα τυχερά παιχνίδια ξεκίνησε η μαθηματική τεκμηρίωση της Θεωρίας Πιθανοτήτων *(Probability Theory)*. Δύο Γάλλοι μαθηματικοί ο Blaise Pascal και ο Pierre de Fermat συνεργάστηκαν μετά από ένα στοίχημα. Ως αποτέλεσμα της συνεργασίας αυτής ήρθε η ανακάλυψη της Θεωρίας των Πιθανοτήτων όπου μέχρι σήμερα αποτελεί τον μαθηματικό πυρήνα της Θεωρίας της Πληροφορίας.

Η μελέτη ζητημάτων που σχετίζονται με τη Θεωρία της Πληροφορίας *(Information Theory)* απαιτούν γνώση βασικών στατιστικών στοιχείων της Θεωρίας Πιθανοτήτων προκειμένου να οριστούν σχετιζόμενες έννοιες όπως το μέτρο ποσότητας πληροφορίας του Hartley αλλά και η μέση ποσότητα πληροφορίας (ή εντροπία) του Shannon. Η ταύτιση των δύο θεωριών είναι εύκολο να διακριθεί εάν αναλογιστεί κανείς πως το σύνολο επιλογής διαθέσιμων συμβόλων στη πηγή ενός προγραμματιστικού περιβάλλοντος αποτελεί το δειγματικό χώρο του πειράματος ενώ αντίστοιχα, το ατομικό αδιαίρετο αποτέλεσμα είναι ένα τυχαίο σύμβολο. Επομένως, στη περίπτωση της ρίψης του ζαριού, ο δειγματικός χώρος του εκάστοτε συμβόλου κατά το σχηματισμό μηνύματος από πηγή πληροφορίας είναι το αλφάβητο που χρησιμοποιείται.

Στο συγκεκριμένο κεφάλαιο θα διευκρινιστούν τα βασικότερα στατιστικά στοιχεία που διαμορφώνουν τη Θεωρία Πιθανοτήτων όπως είναι ο δειγματικός χώρος, η δειγματοληψία, οι διακριτές ή συνεχείς τυχαίες μεταβλητές καθώς και οι συναρτήσεις κατανομής πιθανοτήτων. Επιπλέον, θα γίνει ειδικότερη προσέγγιση τόσο του μέτρου ποσότητας πληροφορίας αναφέροντας τη μαθηματική του τεκμηρίωση αλλά και τη μονάδα μέτρησης του, όσο και της μέσης ποσότητας πληροφορίας περιγράφοντας τις ιδιότητες της αλλά και τις κατηγορίες πληροφορίας.

2.2 Θεωρία Πιθανοτήτων

Η θεωρία πιθανοτήτων αφορά το κομμάτι εκείνο του μαθηματικού κλάδου που ασχολείται με την ανάλυση τυχαίων φαινομένων. Πρόκειται για ένα πιθανολογικό μέτρο και επομένως, κεντρικό ρόλο παίζει η έννοια της πιθανότητας, δηλαδή του αβέβαιου αυτού ποσοστού (θετικός αριθμός, μικρότερος ή ίσος του ένα) που θα επέλθει ένα αποτέλεσμα. Σημαντικές συνιστώσες στη θεωρία αποτελούν στατιστικές έννοιες όπως είναι η δειγματοληψία, ο δειγματικός χώρος μέσα στο οποίο γίνεται ένα πείραμα, ο πληθυσμός και το δείγμα, καθώς και η επιλογή τυχαίων μεταβλητών. Στη συνέχεια, γίνεται μια λεπτομερής ανάλυση των προαναφερόμενων εννοιών προκειμένου να γίνει καλύτερα κατανοητή η διαμόρφωση συναρτήσεων πιθανότητας, η διάκριση στοχαστικών διαδικασιών και τέλος, η Θεωρία Πιθανοτήτων.

Σε ένα σύνολο πλήθους στοιχείων όπως είναι για παράδειγμα ένας πληθυσμός σε ένα χωριό στην Ήπειρο, η επιλογή συγκεκριμένου αριθμού ανθρώπων από εκείνο το χωριό καλείται δειγματοληψία *(Ω)*. Αυτοί οι επιλεγμένοι άνθρωποι θεωρούνται το δείγμα για το πείραμα και ο αριθμός του συνόλου τους συμβολίζεται με *ν* δειγματικά σημεία, ενώ αντίθετα ο συνολικός πληθυσμός του χωριού συμβολίζεται με *N*.

Αυτός ο αριθμός επιλεγμένων ανθρώπων όπου αποτελεί το δείγμα του συνόλου των κατοίκων σε εκείνο το χωριό της Ηπείρου, θεωρείται ο δειγματικός χώρος στο πείραμα τύχης που θα επέλθει. Ως πείραμα τύχης νοείται μια προκαθορισμένη διαδικασία της οποίας το αποτέλεσμα δεν είναι γνωστό εκ των προτέρων, όπως είναι η ρίψη ενός ζαριού ή ενός νομίσματος.

Ο σχηματισμός ενός μέρους του δειγματικού χώρου, γνωστός ως υποσύνολο, αποτελεί ένα γεγονός *(α)*. Με άλλα λόγια, ένα γεγονός είναι μια συλλογή εκβάσεων ή δειγματικών σημείων ή απλών ενδεχομένων του δειγματικού χώρου. Το βέβαιο γεγονός είναι το σύνολο του δειγματοχώρου εφόσον θα συμβαίνει πάντα. Αντίθετα, το μηδενικό γεγονός είναι το υποσύνολο που δεν περιέχει κανένα αποτέλεσμα και συμβολίζεται με $\alpha\{\emptyset\}$.

Αν θεωρηθεί ότι ένα γεγονός α αποτελείται από ν δειγματικά σημεία ενός συνολικού πλήθους N χωρίς συντελεστή βαρύτητας, δηλαδή ισοπίθανα μεταξύ τους, ο ορισμός της πιθανότητας του α θα είναι ο λόγος ν/N, θα συμβολίζεται με $P(\alpha)$ ενώ ταυτόχρονα, θα ικανοποιεί τα εξής αξιώματα της θεωρίας πιθανοτήτων:

1.	$P(\emptyset) = 0$, όπου $\emptyset$ το κενό σύνολο
2.	$P(\Omega) = 1$ *για δειγματικό χώρο* Ω
3.	$0 \leq P(\alpha) \leq 1$ *για γεγονός* $\alpha \in \Omega$
4.	Αν $\alpha_1 \subset \alpha_2$ τότε ισχύει: $$P(\alpha_1) \leq P(\alpha_2)$$
5.	$P(\alpha_2 - \alpha_1) = P(\alpha_2) - P(\alpha_1)$
6.	$P(\alpha^c) = 1 - P(\alpha)$, όπου α^c είναι το συμπλήρωμα του γεγονός α
7.	Για κάθε δύο αποκλειστικά αμοιβαία γεγονότα α_1 και α_2 γνωστά ως $(\alpha_1 \cup \alpha_2)$ ισχύει: $$P(\alpha_1 \cup \alpha_2) = P(\alpha_1) + P(\alpha_2)$$
8.	$P(\alpha \cup \beta) = P(\alpha) + P(\beta) - P(\alpha \cap \beta)$
9.	Αν $\alpha_1, \alpha_2, \ldots, \alpha_N$ είναι γεγονότα ανά δύο αμοιβαία αποκλειστικά τότε ισχύει: $$P\left(\bigcup_{k=1}^{N} \alpha_k\right) = \sum_{k=1}^{N} P(\alpha_k) \ \gamma\iota\alpha\ N \geq 2$$

Η τυχαία επιλογή μιας ή περισσοτέρων στοιχείων του δειγματικού χώρου θεωρείται μια τυχαία και αμερόληπτη διαδικασία, η οποία καλείται πείραμα τύχης. Το επιλεγμένο αυτό στοιχείο ονομάζεται τυχαία μεταβλητή *(X)*. Εναλλακτικά, μια τυχαία μεταβλητή μπορεί να πάρει ένα σύνολο δυνατών τιμών (Ω), σε κάθε μία από τις οποίες αντιστοιχεί μια πιθανότητα (για διακριτές τυχαίες μεταβλητές) ή μια πυκνότητα πιθανότητας (για συνεχείς τυχαίες μεταβλητές). Μέσω μιας τυχαίας μεταβλητής μπορούμε να ορίζουμε και γεγονότα ή ενδεχόμενα. Για παράδειγμα, μια τυχαία μεταβλητή $(X = x_i)$ με δειγματικό χώρο Ω που λαμβάνει πραγματικές τιμές είναι μια απεικόνιση $X : \Omega \in \mathbb{R}$, δηλαδή για κάθε γεγονός $\alpha \in \Omega$ αντιστοιχεί μια τιμή X που $\alpha \in \mathbb{R}$. Σε αυτό το σημείο θα ήταν χρήσιμο να διευκρινιστούν οι έννοιες διακριτή και συνεχής τυχαία μεταβλητή:

- **Διακριτές Τυχαίες Μεταβλητές**

Διακριτή ονομάζεται εκείνη η τυχαία μεταβλητή που μπορεί να πάρει διακριτές τιμές με πεπερασμένο ή αριθμήσιμο πλήθος δυνατών τιμών. Επεξηγηματικά, σε ένα δειγματικό χώρο $\Omega = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ όπου τα $\alpha_1, \alpha_2, \ldots, \alpha_n$ είναι τα διαφορετικά γεγονότα και η διακριτή τυχαία μεταβλητή X έχει πεδίο τιμών $\{x_1, x_2, \ldots, x_n\}$, κάθε γεγονός α_i έχει πιθανότητα να επιλεχθεί ίση με $P(\Omega = \alpha_i) = P(X = x_i) = p_i$. Η εν λόγο πιθανότητες $P(X = x_i)$ διαμορφώνουν την συνάρτηση πιθανότητας όπου καλείται και συνάρτηση μάζας πιθανότητας και ορίζεται ως $f(x_i) : R \rightarrow [0, 1]$ όπου μαθηματικά ισχύει:

$$f(x_i) = P(X = x_i)$$

Θεωρείτε μια πραγματική συνάρτηση και δείχνει την πιθανότητα η τυχαία μεταβλητή X να λάβει μια οποιαδήποτε τιμή $x_i \in$ R. Η κάθε ξεχωριστή πιθανότητα είναι πάντα θετικός αριθμός και επομένως ισχύει:

$$f(x_i) = p_i \geq 0 \; \gamma\iota\alpha \; \kappa\acute{\alpha}\theta\varepsilon \; i$$

Ενώ παράλληλα, αθροίζοντας το σύνολο των πιθανοτήτων ισούται με την μονάδα ή ποσοστιαία με το εκατό. Άρα:

$$\sum_{i}^{n} f(x_i) = 1$$

Επιπλέον, η αναμενόμενη τιμή (ή μέση τιμή) της συνάρτησης δίνεται από την άθροιση του πολλαπλασιασμού του κάθε x_i επι της πιθανότητας η τυχαία μεταβλητή (X) να λάβει αυτό το x_i. Άρα:

$$E(x_i) = \sum x_i P(X = x_i)$$

Στην περίπτωση που η τυχαία μεταβλητή είναι διακριτή, η αθροιστική συνάρτηση κατανομής πιθανότητας θα είναι ασυνεχής και τα σημεία ασυνέχειας θα αντιστοιχούν στις τιμές της τυχαίας μεταβλητής που έχουν θετική πιθανότητα. Η συνάρτηση μάζας πιθανότητας σχετίζεται με την αθροιστική συνάρτηση κατανομής πιθανότητας προσθέτοντας όλες τις προηγούμενες τιμές (x_i) της x και επομένως, ισχύει:

$$f(X \leq x) = \sum_{x_i \leq x} f(x_i) \text{ για κάθε } x \epsilon (-\infty, +\infty)$$

- **Συνεχείς Τυχαίες Μεταβλητές**

Συνεχείς τυχαίες μεταβλητές καλούνται εκείνες οι τυχαίες μεταβλητές (X) που έχουν ως πεδίο τιμών ένα διάστημα του R ή όλο το R. Οι συνεχείς μεταβλητές αντιστοιχούν σε συνεχείς δειγματικούς χώρους. Στην περίπτωση των συνεχών τυχαίων μεταβλητών η έκφραση «η πιθανότητα η μεταβλητή X να πάρει μια ορισμένη τιμή x_i», δηλαδή $P(X = x_i)$, αντικαθίσταται από την πιθανότητα η μεταβλητή X να πάρει τιμές σε ορισμένο διάστημα γύρω από ένα σημείο x_i, δηλαδή να πάρει τιμές γύρω από έναν αριθμό $m_i\ (= x_i)$ όπου είναι το μέσω ενός διαστήματος έστω $[p, q]$. Επομένως, η συνάρτηση κατανομής μιας συνεχούς τυχαίας μεταβλητής δίνεται από το εμβαδόν:

$$P(\alpha \leq x \leq b) = \int_{\alpha}^{b} f(x)\, dx = 1$$

Όπου τα a, b αποτελούν το συνολικό διάστημα μέσα στο όποια υπάρχουν τα ξεχωριστά $m_i\ (= x_i = x)$ ή το δειγματοχώρο που συνήθως είναι $\Omega = R$, δηλαδή $[-\infty, +\infty]$.

Αντίστοιχα, η συνάρτηση αθροιστικής πιθανότητας μιας συνεχούς τυχαίας μεταβλητής παίρνει τη μορφή:

$$f(X \leq x) = P[X \in (-\infty, x)] = \int_{-\infty}^{x} f(y)\, dy \text{ , για κάθε } x \in (-\infty, \infty)$$

όπου $f(y)$, το μέρος της συνάρτησης της αθροιστικής πυκνότητας που ανταποκρίνεται στο διάστημα $X \in (-\infty, x)$.

Η συνάρτηση κατανομής της συνεχούς τυχαίας μεταβλητής έχει τις ακόλουθες ιδιότητες:

$$0 \leq f(X \leq x) \leq 1\ για\ κάθε\ x$$

Δηλαδή, η πιθανότητα στη συνάρτηση f, η τυχαίας μεταβλητή X να λάβει τιμή μικρότερη από το x θα είναι θετικός αριθμός αλλά μικρότερος του ένα. Επιπλέον, η συνάρτηση κατανομής είναι μη φθίνουσα, δηλαδή:

$$Άν\ x_i \leq x\ τότε\ f(X\ \leq x_i) \leq f(X \leq x)$$

Με άλλα λόγια, εάν το x_i είναι μικρότερος αριθμός από την αναμενόμενη τιμή x, τότε το μέρος της συνάρτησης όπου η τυχαία μεταβλητή X είναι μικρότερη από την εκάστοτε x_i είναι σίγουρα σε χαμηλότερο σημείο πάνω στη συνάρτηση από όταν η

τυχαία μεταβλητή είναι μικρότερη από την αναμενόμενη τιμή x. Ενώ ταυτόχρονα, ισχύει η ιδιότητα,

$$lim_{x\to\infty} f(X \leq x) = 1 \; και \; lim_{x\to\infty} f(X \leq x) = 0$$

όπου εκφράζει τα όρια μέσα στα οποία εκτείνεται η συνάρτηση.

2.2.1 Υπό συνθήκη, Συνδυασμένη και Οριακή (ή Ακραία) Πιθανότητα

Πολλές φορές είναι αναγκαίο να συνδυαστούνε n τυχαία πειράματα ή (τυχαίες μεταβλητές) σε ένα ενιαίο. Έστω δύο πειράματα τύχης με ξεχωριστούς δειγματικούς χώρους $X = \{x_1, x_2, \dots, x_n\}$ και $Y = \{y_1, y_2 \dots, y_m\}$ αντίστοιχα. Η συνάρτηση κατανομής πιθανότητας της Y προέρχεται από τις εκάστοτε πιθανότητες. Άρα:

$$\text{Για } P(Y) = \{p_{(y_1)}, p_{(y_2)}, \dots, p_{(y_m)}\} \text{ θα διαμορφωθεί } p_{(y_j)} = P(Y = y_j)$$

Κατόπιν, θα εξεταστεί ένα συνδυαστικό πείραμα τύχης με δειγματικό χώρο το σύνολο των συνδυασμών (X, Y). Ως αποτέλεσμα, θα οριστεί η συνάρτηση συνδυασμένης πιθανότητας μάζας με $p_{ij} = P(X = x_i, Y = y_j)$ που δίνει την πιθανότητα να ισχύει ότι $X = x_i$ και $Y = y_j$. Ως συνέχεια της συνάρτησης θα υπολογιστούν οι συναρτήσεις ακραίας πιθανότητας μάζας $p_{(x_i)}$ και $p_{(y_j)}$ ως εξής :

$$p_{(x_i)} = \sum_{j=1}^{m} p_{ij} \text{ και } p_{(y_j)} = \sum_{i=1}^{n} p_{ij}$$

Στην περίπτωση που το αποτέλεσμα ενός πειράματος Y αποτελεί την συνθήκη για ένα άλλο πείραμα X τότε καλείται υπό συνθήκη πιθανότητα. Χαρακτηριστικό παράδειγμα αποτελεί η υψηλή πιθανότητα εμφάνισης του συμβόλου «η» κατά την λήψη μηνύματος στην ελληνική γλώσσα όταν ο παραλήπτης έχει λάβει ήδη το τμήμα «κυριακ» αφού το επόμενο γράμμα μπορεί να είναι «η» ή «ε». Η συνάρτηση υπό συνθήκη πιθανότητας μάζας $p_{(x_i/y_j)}$ που δίνει την πιθανότητα $X = x_i$ δεδομένου του $Y = y_i$, ορίζεται ακολούθως:

$$p(x_i \,/\, y_j) = \frac{p(x_i, y_j)}{p(y_j)} \text{ εφόσον } p(y_j) > 0$$

Αντίστοιχα, για δεδομένο $X = x_i$, η συνάρτηση υπό συνθήκη πιθανότητας μάζας $p_{(y_j/x_i)}$ που δίνει την πιθανότητα $Y = y_i$ δίνετε από την σχέση:

$$p(y_j \,/\, x_i) = \frac{p(x_i, y_j)}{p(x_i)}, \text{ εφόσον } p(x_i) > 0$$

Ενώ παράλληλα ισχύει η σχέση:

$$\sum_{i=1}^{n} p(x_i \,/\, y_j) = 1$$

Από τα παραπάνω προκύπτει η συνάρτηση συνδυασμένης πιθανότητας μάζας η οποία αναφέρθηκε προηγουμένως πως μας δίνει τις πιθανότητες $X = x_i$ και $Y = y_i$. Άρα:

$$p(x_i, y_j) = p(x_i \,/\, y_j)p(y_j) = P(y_j \,/\, x_i)p(x_i)$$

2.2.2 Ανεξαρτησία Γεγονότων - Θεώρημα Bayes

Γενικότερα, ανεξάρτητα καλούνται δύο γεγονότα, καταστάσεις ή τμήματα που η πραγματοποίηση ή αλλαγή του ενός δεν έχει καμία επίδραση στο άλλο. Με άλλα λόγια, δεν υπάρχει καμία απολύτως συσχέτιση μεταξύ των δύο διαφορετικών γεγονότων. Επομένως, δύο τυχαία γεγονότα A και B ονομάζονται στατιστικά ανεξάρτητα αν ισχύει $P(A \cap B) = P(A)P(B)$. Επεξηγηματικά, η γνώση για την πραγματοποίηση του γεγονότος B δεν παρέχει καμία επιπλέον πληροφορία για την πραγματοποίηση ή μη του γεγονότος A. Σύμφωνα με τα παραπάνω ισχύει:

$$P(A/B) = P(A) \text{ και } P(B/A) = P(B)$$

Τα ανεξάρτητα πειράματα και ειδικότερα οι επαναλαμβανόμενες ανεξάρτητες δοκιμές, βρίσκονται στον πυρήνα πολλών πιθανοθεωρητικών μοντέλων και (γι' αυτό) κατέχουν σημαντική θέση στη Θεωρία Πιθανοτήτων.

Το πιθανοθεωριτικό θεώρημα του Bayes βασίζεται στην επέκταση των δύο εννοιών, της δεσμευμένης πιθανότητας και του θεωρήματος της ολικής πιθανότητας. Με τον όρο δεσμευμένη πιθανότητα ή υπό συνθήκη πιθανότητα νοείται εκείνη η εκ των υστέρων πιθανότητα που διαμορφώνεται για ένα γεγονός, έστω A, μόνο και μόνο όταν έχει πραγματοποιηθεί προηγουμένως ένα άλλο ανεξάρτητο γεγονός, έστω B. Επομένως, εάν έχω δύο ενδεχόμενα A και B του δειγματικού χώρου Ω ενός πειράματος τύχης με $P(B) > 0$, τότε η δεσμευμένη πιθανότητα του A δοθέντος του B συμβολίζεται με $P(A|B)$ και δίνετε από τον τύπο :

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

Ομοίως, αν $P(A) > 0$, τότε:

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

Από τις δύο αναφερόμενες σχέσεις, προκύπτει η συνάρτηση συνδυασμένης πιθανότητας μάζας υπό τη μορφή:

$$P(A \cap B) = P(A \,/\, B)P(B) = P(B \,/\, A)P(A) \qquad 2.1$$

Ενώ παράλληλα, με την υπό συνθήκη πιθανότητα μάζας ισχύει και η σχέση:

$$\sum_{i,j=1}^{n,m} P\left(A_i|B_j\right) = 1$$

Όταν δίνεται η υπό συνθήκη πιθανότητα μάζας $P\left(B_j|A_i\right)$ και η αντίστοιχη $P(A_i)$ είναι πολύ εύκολο μέσω του δοθέντος θεωρήματος του Bayes να βρεθεί το $P\left(A_i \,/\, B_j\right)$. Επομένως, εάν ισχύει η Εξίσωση 2.1 και $P\left(B_j\right) > 0$, μέσω του θεωρήματος της ολικής πιθανότητας προσδιορίζεται το $P\left(A_i \,/\, B_j\right)$ ως εξής:

$$P\left(A_i \,/\, B_j\right) = \frac{P\left(B_j|A_i\right)P(A_i)}{P\left(B_j\right)} = \frac{P\left(B_j \,/\, A_i\right)P(A_i)}{\sum_{i,j=1}^{n,m} P(A_i)P\left(B_j|A_i\right)}$$

Ο κανόνας του Bayes είναι ένα πολύ σημαντικό θεώρημα για την εκτίμηση προβλημάτων στα σήματα.

2.3 Το Μέτρο Πληροφορίας του Hartley

Το έργο του Hartley είχε βάσεις επιρροής μέσα από τον νόμο που σχεδόν ταυτόχρονα ανακάλυψαν το 1924 ο Nyquist στις Ηνωμένες Πολιτείες της Αερικής και ο Kupfmuller στη Γερμανία. Σύμφωνα με εκείνους, η μετάδοση σημάτων τηλεγράφου σε ένα δεδομένο ρυθμό απαιτεί ένα καθορισμένο εύρος συχνοτήτων. Μεταξύ εκείνων που συνέβαλαν στην επαλήθευση του νόμου ήταν ο D. Gabor το 1946 και ο D. M. Mackay το 1948. Χάρη στα έργα διαμόρφωσης και επαλήθευσης από τους προαναφερόμενους, ο νόμος αυτός αποτέλεσε βασικό συστατικό στην μεταγενέστερη ανάπτυξη της Θεωρίας Πληροφορίας.

Ο Hartley διαπίστωσε πως η πληροφορία μπορεί να έχει ποσοτική υπόσταση και να μην ταυτίζεται καθόλου με το σημασιολογικό χαρακτήρα που της προσδίδεται συχνά στην καθημερινή χρήση. Το 1928, ο θεωρητικός της πληροφορίας Ralph V. R. Hartley δημοσίευσε το έργο του με τίτλο *«Μετάδοση Πληροφοριών - Transmission of*

Information», στο οποίο απέδειξε μαθηματικά ότι η συνολική ποσότητα πληροφοριών που μπορεί να μεταδοθεί είναι ανάλογη με το εύρος συχνοτήτων που εκπέμπεται και τον χρόνο μετάδοσης. Με άλλα λόγια, αντιλήφθηκε πως όσο πιο μεγάλη είναι η πιθανότητα εμφάνισης ενός γεγονότος τόσο πιο μικρή είναι η αβεβαιότητα για το αν θα συμβεί. Δηλαδή, στην περίπτωση που αυτό το συχνό γεγονός συμβεί, η πληροφορία που λαμβάνεται είναι μικρή ενώ εναλλακτικά, θα είναι περισσότερη. Για παράδειγμα, εάν δημοσιευθεί στις εφημερίδες πως η Αμερική εισβάλει στο Ιράκ θα είναι ένα γεγονός που δεν θα ξαφνιάσει κανέναν οπότε θα προσφέρει μικρή πληροφορία. Αντιθέτως εάν δημοσιευθεί το αντίθετο θα θεωρηθεί ένα πρωτοφανές γεγονός το οποίο θα παρέχει μεγαλύτερη πληροφορία. Αυτό συμβαίνει επειδή το δεύτερο γεγονός είναι πιο σπάνιο και έχει μικρή πιθανότητα να συμβεί σε σχέση με το πρώτο που είναι σχεδόν βέβαιο.

Έστω A ένα τυχαίο γεγονός με πιθανότητα $P(A)$ και $I(A)$ είναι η συνάρτηση του μέτρου της πληροφορίας του A, τότε η $I(A)$ θα πρέπει να ικανοποιεί τις παρακάτω ιδιότητες:

1. Όταν η πιθανότητα να συμβεί ένα γεγονός είναι ίση με τη μονάδα, δηλαδή δεδομένο (100%), τότε είναι λογικό η ποσότητα της μεταφερόμενης πληροφορίας να είναι μηδενική, δηλαδή σχεδόν ανούσια. Επομένως:
$$Εάν\ P(A) = 1\ τότε\ I(A) = 0$$
2. Η συνάρτηση του μέτρου της πληροφορίας του A είναι θετική καθώς η πιθανότητα ενός γεγονότος να συμβεί είναι πάντα ένα μη αρνητικό μέγεθος, δηλαδή:
$$I(A) \geq 0 \quad διότι\ \ 0 \leq P(A)\ \leq 1$$
3. Όσο πιο σπάνιο το φαινόμενο να πραγματοποιηθεί ένα γεγονός, τόσο μεγαλύτερη πληροφορία λαμβάνεται. Άρα, για ένα γεγονός A όπου έχει μεγαλύτερη πιθανότητα να εμφανιστεί από το B, ισχύει:
$$Για \quad P(B)\ \leq P(A) \quad τότε \ \ I(A) \leq I(B)$$
4. Τέλος, αν τα δύο αναφερόμενα γεγονότα A και B είναι ανεξάρτητα μεταξύ τους με αντίστοιχες πιθανότητες $P(A)$ και $P(B)$, τότε το μέτρο της πληροφορίας του γεγονότος εμφάνισης και των δύο επιμέρους γεγονότων είναι ίσο με το άθροισμα των δύο επιμέρους μέτρων πληροφορίας. Άρα:
$$Άν \ \ P(AB) = P(A)P(B) \ \ τότε$$
$$I(AB) = -\log_b(P_A P_B) \Rightarrow$$
$$I(AB) == -\log_b(P_A) - \log_b(P_A) \Rightarrow$$

$$I(AB) = I(A) + I(B)$$

Με βάση τα παραπάνω ορίστηκε πως η συνάρτηση πληροφορίας ως $I(A)$, η οποία αποκτάται από την πραγματοποίηση ενός γεγονότος A με πιθανότητα εμφάνισης (P_A), δίνεται ως εξής:

$$I(A) = -\log_b P_A = \log_b \left(\frac{1}{P_A}\right),$$

όπου b αποτελεί τη βάση του λογάριθμου και μπορεί ο βαθμός του να επιλεγεί ελεύθερα αρκεί $b > 1$. Η μονάδα μέτρησης της πληροφορίας καθορίζεται ανάλογα με τη βάση υπολογισμού του λογάριθμου. Ο φυσικός λογάριθμος *(natural)* έχει ως μονάδα μέτρησης το nat ενώ ο δεκαδικός λογάριθμος, όπου χρησιμοποιείται από τον Hartley, έχει ως μονάδα μέτρησης το decit *(decimal unit)* γνωστό και ως μονάδα Hartley. Εξετάζοντας το σχηματισμό μηνυμάτων μήκους ενός συμβόλου από ένα αλφάβητο αποτελούμενο από δέκα σύμβολα, η ποσότητα πληροφορίας κάθε μηνύματος θα είναι ίση με:

$$H(N^1) = \log_{10} 10 = 1\ decit$$

Η επικρατέστερη όμως μονάδα μέτρησης της πληροφορίας είναι το bit *(binary unit)* λόγω του δυαδικού συστήματος αρίθμησης στους υπολογιστές. Για αυτή τη περίπτωση, αρκετές φορές δεν αναφέρεται καν η βάση διότι εννοείται ως μονάδα το bit. Επομένως, εξετάζοντας το σχηματισμό μηνυμάτων μήκους ενός συμβόλου από ένα αλφάβητο αποτελούμενο από δύο σύμβολα, τότε η ποσότητα πληροφορίας θα είναι:

$$H(N^1) = \log_2 N = \log_2 2 = 1\ bit$$

Πολλές φορές απαιτείται μετασχηματισμός από τη μια λογαριθμική βάση στην άλλη. Αυτό είναι εφικτό μέσω της διαμορφωμένης σχέσης:

$$\log_b x = \frac{\log_\alpha x}{\log_\alpha b}$$

Σύμφωνα με την πρόταση του Hartley, η ποσότητα πληροφορίας διαμορφώνεται ως το δεκαδικό λογάριθμο του πλήθους των διαφορετικών λέξεων $(N\nu)$ που μπορούν να σχηματιστούν, αποτελούμενες από ένα δεδομένο πλήθος συμβόλων (N). Στην περίπτωση μηνυμάτων μήκους k συμβόλων από ένα αλφάβητο με N σύμβολα, η ποσότητα πληροφορίας είναι ίση με:

$$H(N^k) = \log(N^k) = k \log N$$

όπου k, το μήκος των συμβόλων. Επομένως, η ποσότητα πληροφορίας ενός μηνύματος αποτελούμενου από k σύμβολα θα πρέπει να είναι k φορές μεγαλύτερη από αυτή ενός μηνύματος που αποτελείται από ένα σύμβολο.

Η σύγχρονη Θεωρία Πληροφορίας με την εισαγωγή των διαφορετικών πιθανοτήτων στον ορισμό της ποσότητας πληροφορίας διαμορφώθηκε μόνο μετά την ολοκλήρωση του έργου του Shannon. Η διαφορά με την προ εισηγμένη έννοια του ορισμού της πληροφορίας από τον Hartley, αφορούσε την ελλιπής διάκριση αυτών των διαφορετικών πιθανοτήτων στα σύμβολα που απαρτίζουν το αλφάβητο, καθώς θεωρεί την επιλογή καθενός εξ' αυτών κατά το σχηματισμό ενός μηνύματος ως ίσης πιθανότητας γεγονός.

2.4 Εντροπία

Για κάθε κατανομή πιθανότητας ορίζεται μια ποσότητα που ονομάζεται εντροπία, η οποία έχει πολλές ιδιότητες που συμφωνούν με όσα διαισθητικά αναμένονται από ένα μέτρο ποσότητας πληροφορίας. Συγκεκριμένα το 1948, ο Shannon εισήγαγε την έννοια της εντροπίας όπου νοείται ένα μέτρο της αβεβαιότητας μιας τυχαίας μεταβλητής. Έστω X μια διακριτή τυχαία μεταβλητή με δειγματικό της χώρο $X = \{x_1, x_{2,\dots}, x_n\}$ και συνάρτηση μάζας πιθανότητας $p(x_i) = P(X = x_i)$. Για δύο $p(x_i)$ και $p(y_j)$, όπου αναφέρονται σε δύο διαφορετικές τυχαίες μεταβλητές, στην πραγματικότητα νοούνται ως δύο διαφορετικές συναρτήσεις μάζας πιθανότητας. Η μέση ποσότητα πληροφορίας της X, όπου συμβολίζεται ως $H(X)$, δεδομένου του ορισμού της πληροφορία, $I(x_i) = -log_b p(x_i)$ με $p(x_i) \in [0,1]$, δίνεται από τη σχέση:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot logp(x_i)$$

Η βάση του λογάριθμου συχνά θεωρείται το δύο *(bits)* και επομένως, η εντροπία μετριέται σε δυφίο. Εναλλακτικά, αν η βάση του λογάριθμου είναι το b, θα συμβολίζεται η εντροπία ως $H_b(X)$, ενώ αν είναι το e τότε η εντροπία μετριέται σε nat. Σημειώνεται, πως η εντροπία δεν εξαρτάται από τις πραγματικές τιμές που λαμβάνει η τυχαία μεταβλητή X, αλλά μόνο από τις πιθανότητες που ταυτίζονται με την κάθε μια από αυτές τις πραγματικές τιμές.

Είναι σημαντικό να αναφερθεί, πως η εντροπία μιας τυχαίας μεταβλητής X μπορεί να ερμηνευτεί ως η αναμενόμενη τιμή της συνάρτησης της τυχαίας μεταβλητής $I(\mathrm{X}) = -log_b p(\mathrm{X}) = log_b\left(\frac{1}{p(\mathrm{X})}\right)$, όπου η X λαμβάνεται σύμφωνα με τη συνάρτηση μάζας πιθανότητας $p(x_i) = P(X = x_i)$. Επομένως, συμβολίζεται ως εξής:

$$E[I(X)] = \sum_{i=1}^{n} I(x_i)p(X = x_i)$$

και συνεπώς, ισχύει:

$$H(X) = E_p \, log \frac{1}{p(X)}$$

όπου με E συμβολίζεται η αναμενόμενη τιμή μιας συνάρτησης, δηλαδή η μέση τιμή της.

Σύμφωνα με όλα τα παραπάνω, διαμορφώνονται συνοπτικά πέντε βασικές ιδιότητες που αφορούν την έννοια της εντροπίας ή εναλλακτικά, της μέσης ποσότητας πληροφορίας. Επομένως,

1. Εντροπία καλείται το μέτρο της μέσης αβεβαιότητας μιας τυχαίας μεταβλητής.
2. Η συνάρτηση που περιγράφει την εντροπία ορίζει τον μέσο αριθμό των $bits$ που απαιτούνται για να περιγράψουν την τυχαία μεταβλητή.
3. Η εντροπία μιας τυχαίας μεταβλητής εξαρτάται μόνο από τις πιθανότητες που την χαρακτηρίζουν.
4. Ισχύει ότι $0\,log(0) = 0$, αφού $\lim_{x \to 0}(x \, log \, x) \to 0$. Επομένως, η προσθήκη όρων μηδενικής πιθανότητας δεν μεταβάλλουν καθόλου την εντροπία, δηλαδή την μέση τιμή της συνάρτησης.
5. Η εντροπία σύμφωνα με τον παραπάνω ορισμό μπορεί να θεωρηθεί συνάρτηση της τυχαίας μεταβλητής και αυτός είναι ο λόγος που διαπιστώνεται η δεύτερη ερμηνεία του όρου ως αναμενόμενη τιμή.

2.4.1 Δυαδική Πηγή

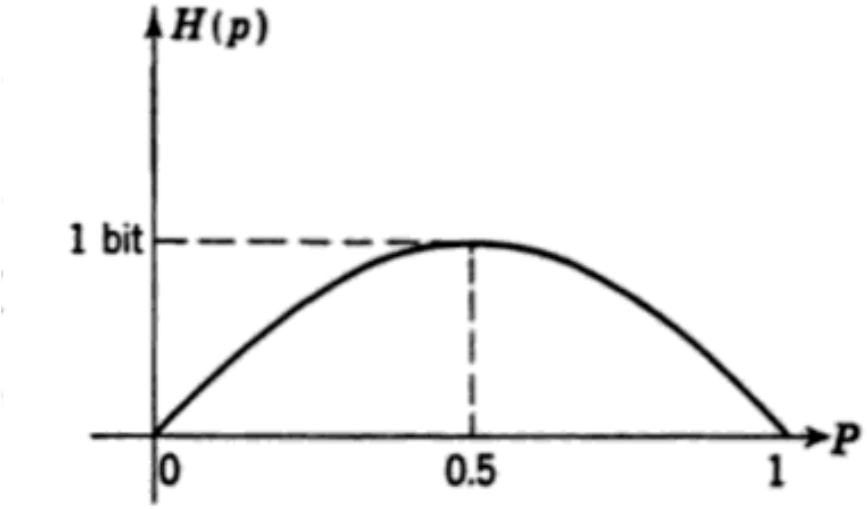
Η δυαδική πηγή *(binary source)* σχετίζεται με εκείνον τον δειγματοχώρο ενός τυχαίου δυαδικού πειράματος όταν το πείραμα επαναλαμβάνεται συνέχεια. Έστω λοιπόν, πως το τυχαίο πείραμα έχει μόνο δύο πιθανά σενάρια επιλογής, οπού θα μπορούσαν να τεθούν ως η πιθανή επιλογή εκπομπής ενός και μόνο μηνύματος από την επιλογή δύο υπαρχόντων, επομένως του A και του B. Στην συνέχεια, αναφέρονται οι βασικές ποσοτικές έννοιες που εντάσσουν την Θεωρία της Πληροφορίας μέσα στη δυαδική πηγή. Άρα:

- ✓ Δειγματοχώρος ή Αλφάβητο $= \{Γράμματα\} = [A, B]$
- ✓ Πιθανότητα Συνολικού Πίνακα $[P] = [p, 1-p] = [p, q]$
- ✓ Ποσότητα Πληροφορίας ενός γεγονότος = $[-\,log\, p\,, -\,log(1-p)]$

Σύμφωνα με τα παραπάνω, τα οποία εκφράζουν την Θεωρία Πληροφορίας, η μέση ποσότητα πληροφορίας ανά γράμμα θα συμβολίζεται με $H(p)$ ή $H(X)$ και μαθηματικά εκφράζεται με τη μορφή:

$$H(p) = -p\,log\,p - q\,log\,q = -p\,log\,p - (1-p)\,log(1-p)$$

Στη γραφική παράσταση του Διαγράμματος 2.1, όπου δείχνει τη συμπεριφορά της μέσης ποσότητας πληροφορίας ως συνάρτηση της πιθανότητας p, γίνεται εφικτή η οπτική κατανόηση της έννοιας.



Πηγή: An Introduction to Information Theory by Fazlollah M. Reza, 1961

Διάγραμμα 2.1 - Η συμπεριφορά της εντροπίας ως συνάρτηση πιθανότητας

Παρατηρείται πως η μέση πληροφορία είναι μια κοίλη συνάρτηση της πιθανότητας και παίρνει τη μέγιστη τιμή $[H(p) = 1bit]$ όταν τα δύο γεγονότα μπορούν να συμβούν με τη ίδια πιθανότητα, δηλαδή στο $p = \frac{1}{2} = 0{,}5$ όπου η εντροπία είναι ίση με $1\ bit$ ανά γράμμα. Επεξηγηματικά, εάν ένας πομπός στείλει τα δύο γράμματα A και B με ίσες πιθανότητες, η μέση πληροφορία ανά γράμμα θα είναι μάξιμουμ $1\ bit$ ανά γράμμα. Αντίθετα, όταν $p = 1$ ή $p = 0$, τότε η εντροπία είναι ίση με $0\ bits$ εφόσον το τελικό αποτέλεσμα είναι βέβαιο.

2.4.2 Ιδιότητες της Μέσης Ποσότητας Πληροφορίας του Shannon

Οι ιδιότητες της μέσης ποσότητας πληροφορίας αφορούν τις βασικές συνιστώσες για τον ορισμό της εν λόγο έννοιας και έχουν τεθεί από τον Shannon μέσω της ερευνάς του για την κατάλληλη συνάρτηση. Εν ολίγης, διακρίνονται στις ακόλουθες τέσσερις ιδιότητες ως εξής:

1. Η μέση πληροφορία $H(p)$ είναι συνεχείς στο p (Διάγραμμα 2.1).
2. Η μέση πληροφορία $H(p)$ είναι συμμετρική, δηλαδή η διάταξη των πιθανοτήτων δεν την επηρεάζει. Έτσι, διαφορετικές τυχαίες μεταβλητές με κατανομές πιθανοτήτων που προέρχονται από μεταθέσεις της ίδιας κατανομής πιθανοτήτων έχουν ίση εντροπία. Υπάρχουν όμως και περιπτώσεις που και διαφορετικές κατανομές πιθανοτήτων οδηγούν στην ίδια μέση ποσότητα πληροφορίας.
3. Η εντροπία $H(p)$ παίρνει τη μέγιστη τιμή όταν όλα τα ενδεχόμενα είναι ισοπίθανα. Σε αυτό το σημείο, η αβεβαιότητα είναι η μέγιστη δυνατή και, κατά συνέπεια, η επιλογή ενός μηνύματος προσφέρει τη μέγιστη δυνατή μέση πληροφορία.
4. Η εντροπία είναι προσθετική. Η ιδιότητα αυτή αναφέρεται στην περίπτωση κατά την οποία δύο ανεξάρτητες τυχαίες μεταβλητές X και Y συνδυάζονται και άρα:

$$H(X,Y) = H(X) + H(Y)$$

Η παραπάνω σχέση είναι εφικτό να δειχθεί μέσω του ορισμού της μέσης πληροφορίας ως εξής:

$$H(X,Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}\, log\, p_{ij}$$

με τα $X = \{x_1, x_2, \dots, x_n\}$ και $Y = \{y_1, y_2, \dots, y_m\}$.

2.5 Συνδυασμένη, Υπό Συνθήκη και Αμοιβαία Πληροφορία

Η εντροπία μπορεί να χρησιμοποιηθεί και για τον ορισμό άλλων μετρήσεων πληροφορίας, οι οποίες αναδεικνύουν τις σχέσεις μεταξύ δύο τυχαίων μεταβλητών X και Y. Έχουμε λοιπόν την από κοινού ή συνδυασμένη ποσότητα πληροφορία(joint entropy), η οποία μετράει τη συνολική πληροφορία των X και Y, την υπό συνθήκη ποσότητα πληροφορίας (conditional entropy), η οποία μετράει την πληροφορία του X, όταν η Y είναι γνωστή και αντίστροφα και τέλος έχουμε την αμοιβαία ποσότητα

πληροφορίας (mutual entropy), η οποία μετράει την σχέση των X και Y, υπό την έννοια ότι μας δείχνει πόσο μειώνεται η πληροφορία του X όταν μαθαίνουμε το Y και αντιστρόφως.

2.5.1 Συνδυασμένη Ποσότητα Πληροφορίας

Υπάρχουν φορές που χρειάζεται να εξεταστεί η ποσότητα πληροφορίας ενός συνδυασμού δύο τυχαίων μεταβλητών (X, Y). Επομένως ένα τυχαίο πείραμα (X, Y) έχει ως δυνατά αποτελέσματα όλους τους συνδυασμούς των αποτελεσμάτων των $X = \{x_1, x_2 \dots x_n\}$ και $Y = \{y_1, y_2 \dots y_m\}$. Δεδομένου αυτού, έχουμε το δειγματοχώρο:

$$\Omega = (X, Y) = \{(x_1, y_1), (x_1, y_2), \dots, (x_1, y_m), \dots, (x_n, y_1), (x_n, y_2), \dots, (x_n, y_m)\}$$

Και η κατανομή πιθανοτήτων δίνεται από:

$$P = \{p(x_1, y_1), \dots, p(x_1, y_m), \dots, p(x_n, y_1), \dots, p(x_n, y_m)$$

Τότε η συνδυασμένη πληροφορία $H(X, Y)$ ορίζεται ως η μέση τιμή:

$$H(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j)\, log\, p(x_i, y_j) \qquad 2.2$$

Οι ακραίες ποσότητες πληροφορίας $H(X)$ και $H(Y)$ μπορούν να υπολογιστούν εφόσον είναι γνωστές όλες οι συνδυασμένες πιθανότητες $p(x_i, y_j)$ καθώς μόνο τότε μπορούν να εκτιμηθούν και οι ακραίες πιθανότητες $p(x_i)$ και $p(y_j)$. Ο ορισμός της μέσης ποσότητας πληροφορίας όπως είναι εκφρασμένος στην εξίσωση 2.2 μπορεί να επεκταθεί και για περισσότερες από δύο διαστάσεις, δηλαδή όταν πρόκειται για περισσότερες από δύο τυχαίες μεταβλητές όπως για παράδειγμα στην περίπτωση $H(X, Y, Z)$ με συνδυασμένη πιθανότητα $p(x_i, y_j, z_k)$, όπου $i = 1,2, \dots, n$, το $j = 1,2, \dots, m$ και τέλος, $k = 1,2, \dots, l$.

Εάν γίνει η υπόθεση ότι όλοι οι δυνατοί συνδυασμοί των τριών τυχαίων μεταβλητών συμβολίζονται με ν και το πλήθος αυτών ισούται με lmn, τότε οι αντίστοιχες πιθανότητες θα οριστούν $p(v_1), p(v_2), \dots p(v_{lmn})$ ενώ η συνδυασμένη πληροφορία δίνεται από την σχέση:

$$H(X, Y, Z) = -\sum_{i=1}^{lmn} p(v_i) \log p(v_i)$$

2.5.2 Υπό Συνθήκη Ποσότητα Πληροφορίας

Η μέση τιμή της υπό συνθήκης πληροφορίας μίας τυχαίας μεταβλητής $X = x_i$ όταν δίνεται το αποτέλεσμα μιας άλλης τυχαίας μεταβλητής $Y = y_j$ διατυπώνεται από τη σχέση:

$$H(x_i \,/\, y_j) = -\log p(x_i \,/\, y_j)$$

Η συγκεκριμένη σχέση θα μπορούσε να διαμορφωθεί για τη μέση τιμή της υπό συνθήκης ποσότητας πληροφορίας μιας τυχαίας μεταβλητής X δεδομένου του αποτελέσματος μιας τιμής y_j ως εξής:

$$H(X \,/\, y_j) = -\sum_{i=1}^{n} p(x_i \,/\, y_j) \log p(x_i, y_j)$$

όπου $X = (x_1, x_2, \dots x_n)$ και $y_j \in Y$.

Η μέση τιμή της υπό συνθήκης ποσότητας πληροφορίας μίας τυχαίας μεταβλητής X δεδομένου όλου του συνόλου των δυνατών αποτελεσμάτων της τυχαίας μεταβλητής Y δίνεται από τον παρακάτω τύπο:

$$H(X \,/\, Y) = \sum_{j=1}^{m} p(y_j) H(X \,/\, y_j) = -\sum_{j=1}^{m} p(y_j) \sum_{i=1}^{n} p(x_i \,/\, y_j) \log(x_i \,/\, y_j)$$

$$= -\sum_{i=1}^{n} \sum_{j=1}^{m} p(y_j) p(x_i \,/\, y_j) \log p(x_i \,/\, y_j)$$

Η φυσικότητα του ορισμού της από κοινού εντροπίας και της υπό συνθήκης εντροπίας αποκαλύπτεται από το γεγονός ότι η εντροπία ενός ζεύγους τυχαίων μεταβλητών $[X, Y]$, είναι η εντροπία της μιας συν την υπό συνθήκη εντροπία της άλλης. Αυτό αποδεικνύεται από τον κανόνα της αλυσίδας.

$$H(X, Y) = H(X) + H(Y|X)$$

Ο εν λόγο κανόνας διατυπώνεται μέσω της παρακάτω απόδειξης:

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\, log\, p(x,y)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\, log\, p(x)p(y|x)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\, log\, p(x)$$

$$-\sum_{x\in X}\sum_{y\in Y} p(x,y)\, log\, p(y|x)$$

$$= -\sum_{x\in X} p(x) log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\, log\, p(y|x)$$

$$= H(X) + H(Y|X)$$

2.5.3 Αμοιβαία Ποσότητα Πληροφορίας

Ως αμοιβαία πληροφορία νοείται το μέτρο της ποσότητας πληροφορίας που περιέχει μια τυχαία μεταβλητή X για κάποια άλλη τυχαία μεταβλητή Y. Με άλλα λόγια, πρόκειται για την μείωση που υφίστανται η αβεβαιότητα μιας τυχαίας μεταβλητής X λόγο της γνώσης της άλλης Y. Έστω δύο τυχαίες μεταβλητές όπου $X = (x_1, x_2, .., x_n)$ και $Y = (y_1, y_2, \dots, y_m)$ είναι δύο πηγές με από κοινού συνάρτηση μάζας πιθανότητας $p(x_i, y_j)$ και με κατανομές πιθανοτήτων $p(x_i)$ και $p(y_j)$. Η αμοιβαία πληροφορία $I(X;Y)$ είναι η σχετική εντροπία μεταξύ της από κοινού κατανομής και της κατανομής γινομένου $p(x_i)\, p(y_j)$ και σχηματίζεται ως εξής:

$$I(X;Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\, log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\, log \frac{p(x_i|y_j)}{p(x_i)}$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\, log\, p(x_i)$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\, log\, p(x_i|y_j)$$

$$= -\sum_{i=1}^{n} p(x_i) \log p(x_i) - \left(-\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log p(x_i|y_j) \right)$$

$$= H(X) - H(X|Y)$$

Ενώ παράλληλα, λόγω συμμετρίας $[I(X;Y) = I(Y;X)]$ έπεται ότι:

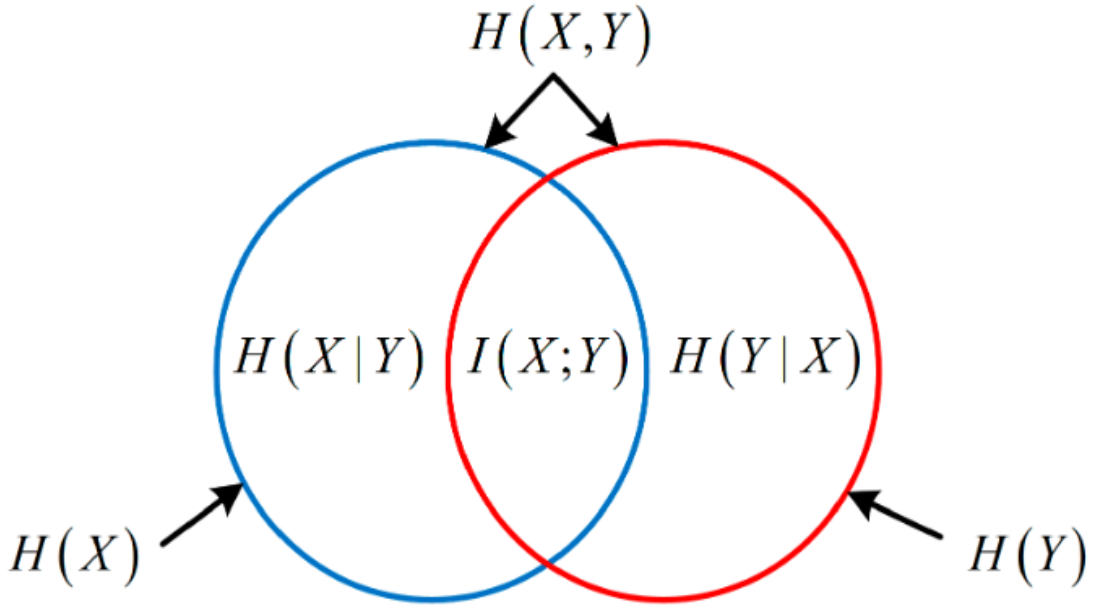
$$I(X;Y) = H(Y) - H(Y|X)$$

Άρα, η X λέει για την Y όσα η Y για την X. Επίσης παρατηρείται ότι:

$$I(X;X) = H(X) - H(X|X) = H(X)$$

Η αμοιβαία πληροφορία μιας τυχαίας μεταβλητής με τον εαυτό της είναι η εντροπία της τυχαίας μεταβλητής. Αυτός είναι ο λόγος για τον οποίο μερικές φορές η εντροπία καλείται αυτοπληροφορία.

Στην Εικόνα 2, διακρίνονται όλες οι περιπτώσεις της εντροπίας, όπως είναι η συνδυασμένη, η υπό συνθήκη και η αμοιβαία ποσότητα πληροφορίας καθώς και οι ακραίες ποσότητες πληροφορίας.



Πηγή: Research Gate, 2011

Εικόνα 2 Συνδυασμένη, Υπό Συνθήκη, Αμοιβαία και Ακραίες Ποσότητες Πληροφορίας

2.6 Ανακεφαλαίωση

Σε αυτό το κεφάλαιο έγινε λεπτομερής ανάλυση της θεωρίας πιθανοτήτων και όλων των σχετιζόμενων εννοιών που την αφορούν, όπως είναι η δειγματοληψία, ο συνολικός πληθυσμός και το δειγματικό πλήθος, το πείραμα τύχης, ένα ενδεχόμενο γεγονός καθώς και οι βασικότερες ιδιότητες της θεωρίας πιθανοτήτων. Επιπλέον, αναφέρθηκε η έννοια της τυχαίας μεταβλητής και των δύο υποκατηγοριών της, γνωστές ως διακριτές και συνεχείς. Κατόπιν, περιγράφηκαν οι τρείς κατηγορίες της πιθανότητας ενώ παράλληλα, ορίστηκε το θεώρημα του bayes βασισμένο σε εκείνες. Στην συνέχεια, διευκρινίστηκε το μέτρο πληροφορίας που όρισε ο Hartley το 1928 πάνω στο οποίο βασίστηκε η θεωρία του Shannon σχετικά με την μέση ποσότητα πληροφορίας. Στο τέλος του κεφαλαίου, αναλύονται οι σημαντικές έννοιες για τη θεωρία του Shannon, δηλαδή η συνδυασμένη, η υπό συνθήκη και η αμοιβαία ποσότητα πληροφορίας.

ΚΕΦΑΛΑΙΟ 3

ΠΗΓΕΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ ΚΩΔΙΚΟΠΟΙΗΣΗΣ

3.1 Εισαγωγή

Η διαδικασία διάδοσης μιας πληροφορίας είναι ευρέως γνωστή και ταυτίζεται με την μετάδοση της από τον πομπό (πηγή) προς τον δέκτη. Σε ένα σύστημα ψηφιακής επικοινωνίας, η εν λόγο διαδικασία απευθύνεται από τον ένα υπολογιστή, ο οποίος δημιουργεί και κωδικοποιεί την πληροφορία προς έναν άλλον υπολογιστή που εκτελεί την αποκωδικοποίηση του μηνύματος για λήψη της πληροφορίας.

Η πηγές πληροφοριών (ή πηγές επικοινωνίας), από τις οποίες διαμορφώνεται το μήνυμα μετάδοσης, μπορούν να ταξινομηθούν ως Διακριτές *(Discrete)* και Συνεχείς *(Continuous)* Πηγές. Στην πρώτη κατηγορία, εντάσσονται οι Διακριτές πηγές χωρίς μνήμη *(Memoryless Discrete Sources-MDS)* και εκείνες με Μνήμη, γνωστές με την μοντελοποιημένη μορφή των Αλυσίδων Markov *(Markov Chains)*. Η δεύτερη κατηγορία, όπου αποτελεί τις Συνεχείς Πηγές αναλύεται στις Γκαουσιανές Πηγές (Gaussian Sources) και στις Μη Γκαουσιανές Πηγές *(non-Gaussian Sources)* ενώ παράλληλα, ταυτίζεται με εκείνες τις πηγές που η έξοδος τους είναι ένα σήμα στον χρόνο.

Σε αυτή την εργασία θα εξεταστεί μόνο η πρώτη κατηγορία των Διακριτών Πηγών. Θα δοθεί ο ορισμός και τα κύρια χαρακτηριστικά των Διακριτών Πηγών χωρίς μνήμη. Παράλληλα, θα περιγραφούν οι βασικές έννοιες που διέπουν τις χωρίς μνήμη πηγές, όπως εκείνες της εντροπίας *(Entropy),* του ρυθμού εντροπίας *(Entropy Rate)* και του πλεονασμού *(Redundancy)* κάθε μηνύματος που μεταδίδεται από μια τέτοια πηγή *(Source)*. Κατόπιν, θα επισημανθούν οι επικρατέστερες μορφές αλγορίθμων *(Algorithm)* που αφορούν την διαδικασία εκτέλεσης της κωδικοποίησης *(Encoding*) για όσο το δυνατόν πιο συμπυκνωμένη αναπαράσταση της πληροφορίας σε κώδικα. Τέλος, θα γίνει αναφορά με παρόμοια διαδικασία ανάλυσης της δεύτερης κατηγορίας των Διακριτών Πηγών με μνήμη. Συγκεκριμένα, θα επεξηγηθούν τα αντίστοιχα γνωστικά αντικείμενα και οι βασικές μαθηματικές έννοιες που διέπουν την συγκεκριμένη κατηγορία.

3.2 Διακριτές Πηγές Πληροφορίας Χωρίς Μνήμη

Ως πηγή *(Source)* σε ένα προγραμματιστικό περιβάλλον *(Programming Environment)* είναι το τμήμα εκείνο σε ένα σύστημα ψηφιακής επικοινωνίας (*Digital Communication System)* που παράγεται η πληροφορία με τη μορφή συμβόλων (γράμματα, αριθμοί, χάρτες, διαγράμματα, κλπ.) μέσα από ένα καθορισμένο αλφάβητο. Συνεπώς, το πλήθος των συμβόλων που υπάρχουν στην πηγή καλείται αλφάβητο πηγής *(Source Alphabet)*.

Η πηγή ταξινομείται σε δύο μεγάλες κατηγορίες, τις Διακριτές *(Discrete)* και τις Συνεχείς πηγές *(Continuous)* αλλά σε αυτή την εργασία θα γίνει αναφορά μόνο στην πρώτη. Μια πηγή από την οποία εκπέμπονται τα δεδομένα σε διαδοχικά διαστήματα και όχι σε συνεχείς, τα οποία είναι ανεξάρτητα από προηγούμενες τιμές χαρακτηρίζεται ως διακριτή πηγή χωρίς μνήμη. Εν ολίγης, αυτή η πηγή ονομάζεται διακριτή λόγω της σταθερής και ανεξάρτητης πιθανότητας επιλογής ενός διαφορετικού κάθε φορά συμβόλου. Αναλυτικά, η διακριτή πηγή παράγει μια σειρά από σύμβολα *(Symbols)* το ένα μετά το άλλο, όπου κάθε σύμβολο (s_i) επιλέγεται από ένα πεπερασμένο αλφάβητο $S = \{s_1, s_2, \dots, s_n\}$ με ρυθμό r_s σύμβολα ανά δευτερόλεπτο. Επισημαίνεται, πως κάθε σύμβολο $s_i \in S$ παράγεται στην πηγή με αμετάβλητη στον χρόνο πιθανότητα ίση με $P(X = s_i) = p_i$.

Μια ομάδα διαδοχικών συμβόλων *(Group of Symbols)* καλείτε λέξη *(Word)*, ενώ μια ομάδα διαδοχικών λέξεων *(Group of Words)* ονομάζεται μήνυμα *(Message)*. Άρα, για πλήθος δυνατών μηνυμάτων από την πηγή ίσο με q, το σύνολο των μηνυμάτων θα συμβολίζεται με $M = \{m_1, m_2, \dots, m_q\}$. Εάν κάθε ένα μήνυμα αποτελείται από l σύμβολα από ένα αλφάβητο συμβόλων $S = \{s_1, s_2, \dots, s_n\}$, τότε το πλήθος των δυνατών μηνυμάτων θα είναι ίσο με $q = n^l$. Επομένως, το μέσο πληροφοριακό περιεχόμενο μηνυμάτων θα δοθεί από την ακόλουθη σχέση:

$$H(M) = -\sum_{i=1}^{q} p(m_i)\, log\, p(m_i)\ \ bits\ /\ message$$

όπου $P = \{p(m_1), p(m_2), \dots, p(m_q)\}$, η κατανομή πιθανοτήτων.

Συνηθίζεται να γίνονται αναφορές σε ομάδες συμβόλων από μια διακριτή πηγή και όχι σε μεμονωμένα τυχαία σύμβολα. Ως ομάδα συμβόλων θεωρούνται k σύμβολα, τα οποία εκπέμπονται από μια διακριτή πηγή σε διαδοχικές χρονικές στιγμές. Ως αποτέλεσμα, αναφορά γίνεται σε ένα νέο αλφάβητο (S^k), το οποίο αποτελείται από n^k

δυνατά σύμβολα, με n το πλήθος των συμβόλων του αλφαβήτου S της αρχικής πηγής. Έτσι, για μια διακριτή πηγή χωρίς μνήμη, όπου τα παραγόμενα σύμβολα είναι στατιστικά ανεξάρτητα μεταξύ τους, η πιθανότητα εμφάνισης ενός συμβόλου από το αλφάβητο S^k είναι ίση με το γινόμενο των πιθανοτήτων της επιμέρους ομάδας k συμβόλων από το αλφάβητο της αρχικής πηγής S. Ισχύει, λοιπόν, η ακόλουθη σχέση που αφορά την εντροπία μιας διακριτής πηγής χωρίς μνήμη και την εντροπία της αντίστοιχης k τάξης:

$$H(S^k) = kH(S)$$

3.2.1 Εντροπία Διακριτής Πηγής Πληροφορίας Χωρίς Μνήμη

Το αποτέλεσμα της διαδικασίας αποστολής μιας πληροφορίας μέσα σε ένα ψηφιακό σύστημα επικοινωνίας είναι, φυσικά, η ορθή μετάδοση ενός ή πολλών μηνυμάτων. Τα μηνύματα, όπως αναφέρθηκε και προηγουμένως, αποτελούνται από επιμέρους λέξεις, δηλαδή διαδοχικές ακολουθίες επιλογής τυχαίων συμβόλων. Ως συνέπεια, ένα επικοινωνιακό σύστημα απασχολείται με την διάκριση των επιμέρους συμβόλων, την πυκνότητα περιεχομένου (μέση ποσότητα περιεχομένου) και τον ρυθμό μετάδοσης πληροφορίας των εν λόγο συμβόλων κάθε μηνύματος.

Η εντροπία ή μέση ποσότητα πληροφορίας μιας διακριτής πηγής χωρίς μνήμη με αλφάβητο $S = \{s_1, s_2, \dots, s_n\}$ όπου n το πλήθος των συμβόλων του αλφαβήτου και p_i η πιθανότητα επιλογής του συμβόλου s_i, δίνεται από την παρακάτω σχέση:

$$H(S) = -\sum_{i=1}^{n} p_i \log p_i \ \ bits/\ symbol$$

Από την παραπάνω εξίσωση παρατηρούμε ότι η τιμή της εντροπίας εξαρτάται μόνο από τις τιμές των πιθανοτήτων p_i των συμβόλων της πηγής.

Όπως διακρίνεται στο Διάγραμμα 2.1 του προηγούμενου κεφαλαίου, όταν όλα τα σύμβολα μιας πηγής πληροφορίας χωρίς μνήμη έχουν ίσες πιθανότητες, τότε η πηγή παρουσιάζει μέγιστη τιμή εντροπίας $maxH(S)$ η οποία δίνετε από την σχέση:

$$maxH(S) = -\sum_{i=1}^{n} \frac{1}{n} log \frac{1}{n} = log\, n \ \ bits/\ symbol$$

Για να μεταδοθεί ένα ορθός αλλά και όσο των δυνατών μικρότερης διάστασης μήνυμα από την πηγή στον δέκτη, απαιτείται ο εντοπισμός της μέγιστης δυνατής αφαίρεσης πλεονασμού περιεχομένου αυτού του μηνύματος. Ως αποτέλεσμα, θα διαμορφωθεί μια έκδοση συμπιεσμένης αναπαράστασης αυτού του μηνύματος με την εισχώρηση των ουσιαστικότερων σημείων του για υψηλότερη ταχύτητα και απόδοση. Εν ολίγης, ο πλεονασμός μιας διακριτής πηγής εκφράζει το ποσοστό άχρηστης πληροφορίας της τρέχουσας κατάστασης από την ιδανική περίπτωση (στην οποία τα σύμβολα της πηγής είναι ισοπίθανα) και δίνεται υπό την εξής μορφή:

$$red = 1 - \frac{H(S)}{maxH(S)} = 1 - \frac{H(S)}{log\, n}$$

Το red ταυτίζεται με τον πλεονασμό *(Redundancy)*, το $H(S)$ και το $maxH(S)$ αποτελούν αντίστοιχα τη μέση και τη μέγιστη ποσότητα πληροφορίας του περιεχομένου με αλφάβητο *(Alphabet)* αποτελούμενο από n σύμβολα. Είναι σημαντικό να σημειωθεί, πως οι τιμές που δύναται να λάβει ο πλεονασμός θα είναι στο διάστημα $[0,1]$ εφόσον η εντροπία (ή μέση ποσότητα) θα είναι πάντα μικρότερη ή ίση της μέγιστης εντροπίας μιας πηγής. Επιπλέον, θα ήταν χρήσιμο να αναφερθεί πως η εμφάνιση πλεονασμού προέρχεται από το γεγονός ότι τα σύμβολα της πηγής δεν είναι ισοπίθανα μεταξύ τους αλλά πιθανό είναι και το γεγονός ύπαρξης πηγής πληροφορίας με μνήμη *(Information Source with Memory)*. Η δεύτερη περίπτωση θα αναλυθεί εκτενέστερα στην συνέχεια του παρόντος κεφαλαίου.

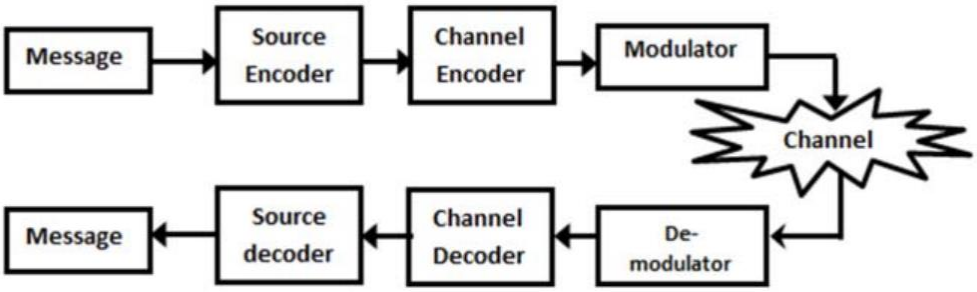
Ενδιαφέρον αποτελεί το γεγονός του ρυθμού μετάδοσης της εν λόγο συμπιεσμένης αναπαράστασης του μηνύματος. Λαμβάνοντας υπόψη, ότι η πηγή παράγει σύμβολα με ρυθμό r_s και παρουσιάζει εντροπία $H(S)$, τότε ο ρυθμός παροχής πληροφορίας από την πηγή R εντοπίζεται από την σχέση:

$$R = r_s \cdot H(S)\ bits\ /\ sec$$

Για παράδειγμα, στη περίπτωση που έχουμε δύο διαφορετικές μεταξύ τους πηγές αλλά με την ίδια εντροπία και ταυτόχρονα, διαφορετικό ρυθμό παροχής πληροφορίας, τότε η πηγή με τον ταχύτερο ρυθμό εκπομπής συμβόλων θα παρέχει προφανώς και περισσότερη πληροφορία.

3.3 Κωδικοποίηση Πηγής

Η ψηφιακή επικοινωνία είναι η μετάδοση σημάτων συμβολικής αξίας από τον πομπό στον δέκτη. Λόγω πολυπλοκότητας και ύπαρξης πλεονασμού, η αποστολή ενός αρχείου ή μηνύματος μέσω του διαδικτύου, πρέπει πρώτα να αναπαρίσταται σε μια αποδοτική πληροφορία μέσω την μοντελοποίησης των αρχικών συμβολών του μηνύματος σε ένα διαφορετικό κώδικα συμβόλων. Με αυτόν τον τρόπο, διαμορφώνεται ένα συμπιεσμένο μήνυμα με την αφαίρεση του προβλήματος που δημιουργεί το πλεοναστικό περιεχόμενο όπου στην συνέχεια, αποστέλλεται στο ψηφιακό κανάλι και εκτελείται η αντίστροφη διαδικασία αποκωδικοποίησης (Εικόνα 3). Συνηθίζεται στο προγραμματιστικό περιβάλλον να χρησιμοποιείται ο δυαδικός κώδικας $[0,1]$, δηλαδή ακολουθίες από $bits$. Ταυτόχρονα, σημαντικό ρόλο στην διαδικασία μετάδοσης παίζει η ταχύτητα μετάδοσης του μηνύματος. Δεδομένου αυτού, επιδιώκεται η χρήση όσο τον δυνατόν λιγότερων $bits$ χωρίς όμως ο δέκτης να μην μπορεί να προσδιορίσει ποιος ήταν ο κάθε χαρακτήρας από την ακολουθία των $bits$ και ως αποτέλεσμα να δημιουργούνται σφάλματα *(Errors)*.



Πηγή: Research Gate, Haldia Institute of Technology

Εικόνα 3 - Διαδικασία Κωδικοποίησης και Αποκωδικοποίησης Μηνύματος

Η εν λόγο διαδικασία καλείται επίσης διαδικασία συμπίεσης. Οι αλγόριθμοι συμπίεσης μετατρέπουν τα αρχικά σύμβολα ενός μηνύματος σε συμπιεσμένη μορφή καθιστώντας το ικανό να ανακτηθεί στην συνέχεια. Τα συστήματα που αφορούν την διαδικασία συμπίεσης και αποσυμπίεσης ταξινομούνται στα στατικά *(Static)* και στα δυναμικά *(Adaptive)* συστήματα. Για την πρώτη κατηγορία, το αρχείο ή μήνυμα παραμένει ως έχει κατά την διαδικασία συμπίεσης και αποσυμπίεσης, ενώ το τελευταίο δύναται να

αλλάζει κατά τη διαδικασία. Σημειώνεται, πως ανάλογα με το μήκος του χρησιμοποιούμενου αλγορίθμου *(Codeword)* κατά τη συμπίεση ή την αποσυμπίεση χαρακτηρίζονται οι ακόλουθες δυνατές κατηγορίες αλγορίθμων:

- **Σταθερό Μήκος**

Για του σταθερού μήκους *(Fixed to Fixed)* αλγόριθμους, η απεικόνιση κάθε συμβόλου αντιστοιχεί σε σταθερό αριθμό $bits$ και κατά τη συμπίεση η κωδικοποίηση γίνεται σαν ακολουθία ενός ορισμένου αριθμού $bits$.

- **Σταθερό προς Μεταβλητό Μήκος**

Η δεύτερη κατηγορία, που αφορά του σταθερού προς μεταβλητού μήκους *(Fixed to Variable)* αλγορίθμους, κάθε σύμβολο πριν τη συμπίεση έχει ορισμένο μήκος αλλά μετά τη συμπίεση μεταβάλλεται το μήκος του.

- **Μεταβλητό προς Σταθερό Μήκος**

Στο μεταβλητό προς σταθερό μήκος *(Variable to Fixed)* αλγόριθμου ισχύει πως ενώ πριν τη συμπίεση μια ακολουθία συμβόλων χαρακτηρίζεται από διαφορετικό αριθμό $bits$ μετά τη συμπίεση κωδικοποιείται σε συγκεκριμένο αριθμό $bits$.

- **Μεταβλητού Μήκους**

Στους αλγόριθμους μεταβλητού μήκους *(Variable to Variable)*, η κωδικοποίηση μιας αλληλουχίας συμβόλων που χαρακτηρίζεται από ένα μεταβλητό αριθμό $bits$ μετά τη συμπίεση μετατρέπεται σε μια διαφορετικού μήκους αλληλουχία συμβόλων.

Σύμφωνα με τη πρωτότυπη διατριβή του Shannon δύο παράμετροι που λαμβάνονται σοβαρά υπόψη κατά την υιοθέτηση ενός σχήματος κωδικοποίησης μιας ψηφιακής πηγής είναι η συχνότητα εμφάνισης *(Frequencies)* και η αυτοσυσχέτιση *(Autocorrelation)* δύο διαδοχικά εμφανιζομένων συμβόλων. Αν οι πιθανότητες εμφάνισης των συμβόλων είναι ίσες και τα διαδοχικά εμφανιζόμενα σύμβολα είναι ανεξάρτητα μεταξύ τους τότε τα σχήματα κωδικοποίησης σταθερού μήκους είναι τα πλέον κατάλληλα. Στην αντίθετη περίπτωση, ο όγκος της μεταβιβαζόμενης κυκλοφορίας μπορεί να μειωθεί αν χρησιμοποιήσουμε σχήματα κωδικοποίησης μεταβλητού μήκους. Σε αυτά τα σχήματα κωδικοποίησης το μήκος της ακολουθίας $bits$ που προσδιορίζει τα σύμβολα είναι μεταβλητό.

Συμπερασματικά, σημαντικό ρόλο παίζει ο αριθμός των $bits$ που χρησιμοποιείται προκειμένου να καθοριστεί τελικά κάθε σύμβολο από το αρχικό αλφάβητο. Έστω πως ένας κώδικας χρησιμοποιεί μ αριθμό $bits$ [0,1], όπου αφορά τον αριθμό των δυνατών συνδυασμών αντιπροσωπευτικών κωδίκων λέξεων, δηλαδή τον αριθμό των αρχικών συμβόλων που μπορεί να περιγράψει με αυτές θα είναι ίσος με 2^{μ}. Σημειώνεται πως τα διαφορετικά κώδικά σύμβολα q (στην προκειμένη περίπτωση $q = 2$) που χρησιμοποιούμε για τη μετατροπή των συμβόλων ή ακολουθιών συμβόλων της πηγής, δηλαδή $S = \{s_1, s_2, \dots, s_n\}$, σε ακολουθίες κωδικών συμβόλων ή κώδικές λέξεις, δηλαδή $C = (c_1, c_2, \dots, c_n)$ απαρτίζουν το επονομαζόμενο κωδικό αλφάβητο. Επομένως, εάν ένας κώδικας επιδιώκει την κωδικοποίηση n αρχικών διαφορετικών συμβόλων, τότε ο αριθμός μ των $bits$ που θα πρέπει να χρησιμοποιήσει δίνεται από τη σχέση:

$$2^{\mu-1} \leq n \leq 2^{\mu}$$

Είναι επιθυμητή ασφαλώς η κατασκευή άμεσων κωδίκων με το ελάχιστο προσδοκώμενο μήκος για τη συμπιεσμένη αναπαράσταση των συμβόλων μιας πηγής. Όπως όμως αναφέρθηκε, δεν γίνεται να δοθούν πολύ μικρές κώδικες λέξεις για όλα τα σύμβολα της πηγής και ο κώδικας να είναι άμεσος (ή προθεματικός), αφού κάποιες θα είναι προθέματα άλλων. Η ανισότητα του Kraft, περιορίζει το σύνολο των δυνατών μηκών των κωδικών λέξεων για να είναι ένας κώδικας άμεσος.

Ανισότητα Kraft

Για κάθε άμεσο κώδικα με πλήθος συμβόλων του κωδικού αλφάβητου q και μήκη των κωδικών λέξεων l_i , όπου $i = 1,2, \dots, n$ και n το πλήθος των κωδίκων συμβόλων της πηγής, ισχύει η ακόλουθη ανισότητα:

$$\sum_{i=1}^{n} q^{-l_i} \leq 1$$

Αντίστροφα, δεδομένου ενός συνόλου μηκών κωδικών λέξεων που ικανοποιούν την ανισότητα, υπάρχει ένας άμεσος κώδικας με κώδικές λέξεις που έχουν αυτά τα μήκη.

Θεώρημα Κωδικοποίησης Πηγής χωρίς μνήμη

Ο Shannon απέδειξε στο μνημειώδες έργο του αυτό που σήμερα καλείται Θεώρημα Κωδικοποίησης Πηγής ή Πρώτο Θεώρημα του Shannon. Σύμφωνα με αυτό, για μια διακριτή πηγή $S = \{s_1, s_2, \dots, s_n\}$ διακριτών συμβόλων με πιθανότητες εμφάνισης ίσες με $p_i = p(X = i)$, οι κώδικές λέξεων που αντιστοιχούν σε κάθε σύμβολο s_i με μήκος l_i, συμβολίζονται με $C = (c_1, c_2, \dots, c_n)$. Άρα, ο μέσου μήκους κώδικας $\overline{B(S)}$, όπου υποδηλώνει τον μέσο αριθμό των $bits$ που χρησιμοποιούνται για την αναπαράσταση ενός συμβόλου s_i, δηλαδή $\overline{B(s_\iota)}$, υπολογίζεται ως:

$$\overline{B(s_\iota)} = \sum_{i=1}^{n} p_i l_i$$

Ο μέσος αριθμός των $bits$ για κάθε σύμβολο απαιτείται προκειμένου να προσδιοριστεί το σύνολο του αλφάβητου που σχηματίζεται ως εξής:

$$\overline{B(S)} = \sum_{i=1}^{n} \overline{B[s_\iota]} p[s_i]$$

Το θεώρημα κωδικοποίησης πηγής για μια πηγή χωρίς μνήμη δηλώνει ότι η εντροπία ενός αλφαβήτου συμβόλων καθορίζει πόσα $bits$ κατά μέσο όρο πρέπει να χρησιμοποιηθούν για την αποστολή ολόκληρου του αλφαβήτου. Επομένως, το θεώρημα δηλώνει ότι ο μέσος αριθμός $bits$ που απαιτούνται για την ακριβή αναπαράσταση του αλφαβήτου πρέπει μόνο να ικανοποιεί την εξής σχέση:

$$H(S) \leq \overline{B(S)}$$

Ταυτόχρονα, για μια δοσμένη διακριτή πηγή χωρίς μνήμη, το μέσο μήκος κώδικα $\overline{B(S)}$, για τους προθεματικούς κώδικες επιδεικνύει πως επαληθεύεται η εξής σχέση με την εντροπία της πηγής των συμβόλων:

$$H(S) \leq \overline{B(S)} \leq H(S) + 1$$

Τονίζεται, πως προθεματικοί κώδικες είναι εκείνοι οι κώδικες που καμία κωδική λέξη δεν αποτελεί πρόθεμα για μια άλλη κωδική λέξη.

Συμπερασματικά, δεν υπάρχει κωδικοποίηση πηγής που πετυχαίνει μικρότερο μέσο μήκος κώδικα από την εντροπία της πηγής αλλά και επιπλέον, να βρίσκεται κοντά σε αυτή. Όσο μικρότερη είναι η εντροπία ενός αλφαβήτου τόσο λιγότερα $bits$ απαιτούνται

για την ψηφιακή μετάδοση μηνυμάτων που εκφράζονται σε αυτό το αλφάβητο. Επισημαίνεται, πως οι αλγόριθμοι κωδικοποίησης, μπορούν να χρησιμοποιηθούν και στις πηγές με μνήμη μέσω της εφαρμογής τους σε μηνύματα ή συνδυασμούς συμβόλων μήκους ίσου με το βάθος της πηγής.

3.3.1 Αλγόριθμοι Κωδικοποίησης

Η λέξη αλγόριθμος *(Algorithm)* προέρχεται από μια μελέτη του Πέρση μαθηματικού Μοχάμεντ Ιμπν Μουσά Αλ Χουαρίζμι (Muhammad ibn Mūsā al-Khwārizmī), που έζησε το 825 μ.Χ. Με τον όρο Αλγόριθμος νοείται μια πεπερασμένη σειρά ενεργειών, αυστηρά καθορισμένων και εκτελέσιμων σε πεπερασμένο χρόνο, που στοχεύουν στην επίλυση ενός προβλήματος. Είναι σημαντικό να διευκρινιστεί πως η έννοια του αλγορίθμου δεν συνδέεται αποκλειστικά και μόνο με προβλήματα της Πληροφορικής αλλά με οποιαδήποτε διαδοχική σειρά αυστηρών ενεργειών που αποσκοπούν στην επίτευξη ενός συγκεκριμένου στόχου.

Από την άλλη πλευρά, η κωδικοποίηση δεδομένων αποτελεί τη μοντελοποίηση των εμπεριεχομένων συμβόλων των δεδομένων σε έναν διαφορετικό κώδικα με σκοπό τη συμπίεση της παρεχόμενης πληροφορίας. Η συμπίεση περιλαμβάνει την εφαρμογή ενός αλγορίθμου στα δεδομένα, ο οποίος καθιστά κάποια από τα επαναλαμβανόμενα $bits$ μη απαραίτητα. Το βασικό χαρακτηριστικό της συμπίεσης είναι ότι όταν τα δεδομένα αποσυμπιεστούν έρχονται στην αρχική τους μορφή.

Εν ολίγης, οι αλγόριθμοι κωδικοποίησης αποτελούν βασικές, αυστηρές επαναλαμβανόμενες, (ψηφιακά) ενέργειες προκειμένου να μοντελοποιηθούν ορισμένα αρχικά σύμβολα σε επιμέρους κώδικά σύμβολα. Με άλλα λόγια, εκτελούνται καθορισμένες κινήσεις αντικατάστασης των αρχικών συμβόλων σε νέα διαφορετικού αλφαβήτου κωδικά σύμβολα. Στην συνέχεια, αναλύονται ορισμένοι από τους επικρατέστερους αλγόριθμους κωδικοποίησης που επιτυγχάνουν την εύρεση αποδοτικών κωδίκων, όπως είναι εκείνοι οι αλγόριθμοι του Fano, του Shannon και του Huffman.

Αλγόριθμος Κωδικοποίησης Fano

Η κωδικοποίηση Shannon–Fano, που πήρε το όνομά της από τους Claude Elwood Shannon και Robert Fano, είναι μια τεχνική για την κατασκευή ενός κώδικα προθέματος που βασίζεται σε ένα σύνολο συμβόλων και των πιθανοτήτων τους. Ο αλγόριθμος λειτουργεί και παράγει αρκετά αποτελεσματικές κωδικοποιήσεις μεταβλητού μήκους. Όταν τα δύο μικρότερα σύνολα που παράγονται από μια κατάτμηση είναι στην πραγματικότητα ίσης πιθανότητας, το ένα bit της πληροφορίας που χρησιμοποιείται για τη διάκρισή τους χρησιμοποιείται πιο αποτελεσματικά.

Ένα δέντρο Shannon–Fano είναι κατασκευασμένο σύμφωνα με μια προδιαγραφή που έχει σχεδιαστεί για να ορίζει έναν αποτελεσματικό πίνακα κωδικών. Ο πραγματικός αλγόριθμος είναι απλός:

1. Για μια δεδομένη λίστα συμβόλων πηγής ή μηνυμάτων, αναπτύξτε μια αντίστοιχη λίστα πιθανοτήτων ή μετρήσεων συχνοτήτων, έτσι ώστε να είναι γνωστή η σχετική συχνότητα εμφάνισης κάθε συμβόλου.
2. Ταξινομήστε τις λίστες συμβόλων ανάλογα με τη συχνότητα (ή πιθανότητα), με τα σύμβολα που εμφανίζονται πιο συχνά στα αριστερά και τα λιγότερο κοινά στα δεξιά.
3. Χωρίστε τη λίστα σε δύο μέρη, με τις συνολικές μετρήσεις συχνοτήτων του αριστερού μέρους να είναι όσο το δυνατόν πιο κοντά στο σύνολο του δεξιού.
4. Χρησιμοποίησε τον δυαδικό κώδικα με n σύμβολα πηγής επιλέγοντας το k έτσι ώστε η ακόλουθη διαφορά των αθροιστικών πιθανοτήτων εμφάνισης των συμβόλων των δύο ομάδων να ελαχιστοποιείται:

$$\left| \sum_{i=1}^{k} p_i - \sum_{i=k+1}^{n} p_i \right|$$

   Στο αριστερό μέρος της λίστας εκχωρείται το δυαδικό ψηφίο 0 και στο δεξί μέρος το ψηφίο 1. Αυτό σημαίνει ότι οι κωδικοί για τα σύμβολα στο πρώτο μέρος θα ξεκινούν όλοι με 0 και οι κωδικοί στο δεύτερο μέρος όλα θα ξεκινήσουν με 1.

5. Εφαρμόστε αναδρομικά τα βήματα 3 και 4 σε καθένα από τα δύο μισά, υποδιαιρώντας ομάδες και προσθέτοντας $bits$ στους κωδικούς έως ότου η κάθε ομάδα να αποτελείται από ένα σύμβολο. Σε κάθε επανάληψη του βήματος 4, επιλέγεται ένα ακόμα κωδικό σύμβολο για το σχηματισμό των κωδικών λέξεων.

Σημειώνεται, πως ο αλγόριθμος του Fano τερματίζει στην περίπτωση πεπερασμένου πλήθους συμβόλων ή μηνυμάτων, αφού η επανάληψη των βημάτων 3 και 4 δεν ξεπερνάει το πλήθος τους. Δυστυχώς, το Shannon–Fano δεν παράγει πάντα βέλτιστους κωδικούς προθέματος. Για το λόγο αυτό, το Shannon–Fano δεν χρησιμοποιείται σχεδόν ποτέ. Η κωδικοποίηση Huffman είναι σχεδόν το ίδιο υπολογιστικά απλή και παράγει κωδικούς προθέματος που επιτυγχάνουν πάντα το χαμηλότερο αναμενόμενο μήκος κωδικών λέξεων. Είναι σημαντικό να τονιστεί πως ο αλγόριθμος Fano μπορεί να οδηγήσει σε άριστους κώδικες αν είναι δυνατή η επαναλαμβανόμενη διαίρεση των ομάδων συμβόλων σε ακριβώς ισοπίθανες (υπό)ομάδες.

Αλγόριθμος Κωδικοποίησης Shannon

Ο αλγόριθμος κωδικοποίησης του Shannon απασχολείται επίσης με την κωδικοποίησης των συμβόλων ή μηνυμάτων της πηγής ενώ παράλληλα, το μήκος της κωδικής λέξης πληροί την συνθήκη ανισότητας Kraft και μπορεί να χρησιμοποιηθεί για τη δημιουργία μοναδικά αποκωδικοποιήσιμων κωδικών. Επομένως, ακολουθεί τα εξής βήματα ανάπτυξης:

1. Τα σύμβολα (ή τα μηνύματα) διατάσσονται σε τάξη φθίνουσας πιθανότητας.
2. Για κάθε σύμβολο s_i του οποίου η πιθανότητα εμφάνισης είναι $p(s_i)$, υπολογίζεται η αθροιστική πιθανότητα P_i, που ορίζεται από τη σχέση:

$$P_i = \sum_{j=1}^{i-1} p(s_j)$$

3. Το πλήθος των κωδικών συμβόλων της κωδικής λέξης η οποία αναπαριστά το σύμβολο της πηγής s_i είναι ίσο με τον ακέραιο αριθμό l_i , που πληροί την ακόλουθη ανισότητα:

$$log\frac{1}{p(s_i)} \leq l_i < 1 + log\frac{1}{p(s_i)}$$

4. Η κωδική λέξη c_i του συμβόλου της πηγής s_i είναι το δυαδικό ανάπτυγμα του κλάσματος P_i (μόνο τα πρώτα l_i $bits$ του αναπτύγματος), δηλαδή ισχύει η σχέση $c_i = (P_i)_{binary\ li\ bits}$. Για παράδειγμα, στο δυαδικό ανάπτυγμα ενός κλάσματος ισχύει:

$$\frac{\alpha_1}{2^1} + \frac{\alpha_2^2}{2^2} + \cdots + \frac{\alpha_k^k}{2^k} = \alpha_1\alpha_2 \ldots \alpha_k \text{ όπου } \alpha_j \text{ είναι } 0 \text{ ή } 1$$

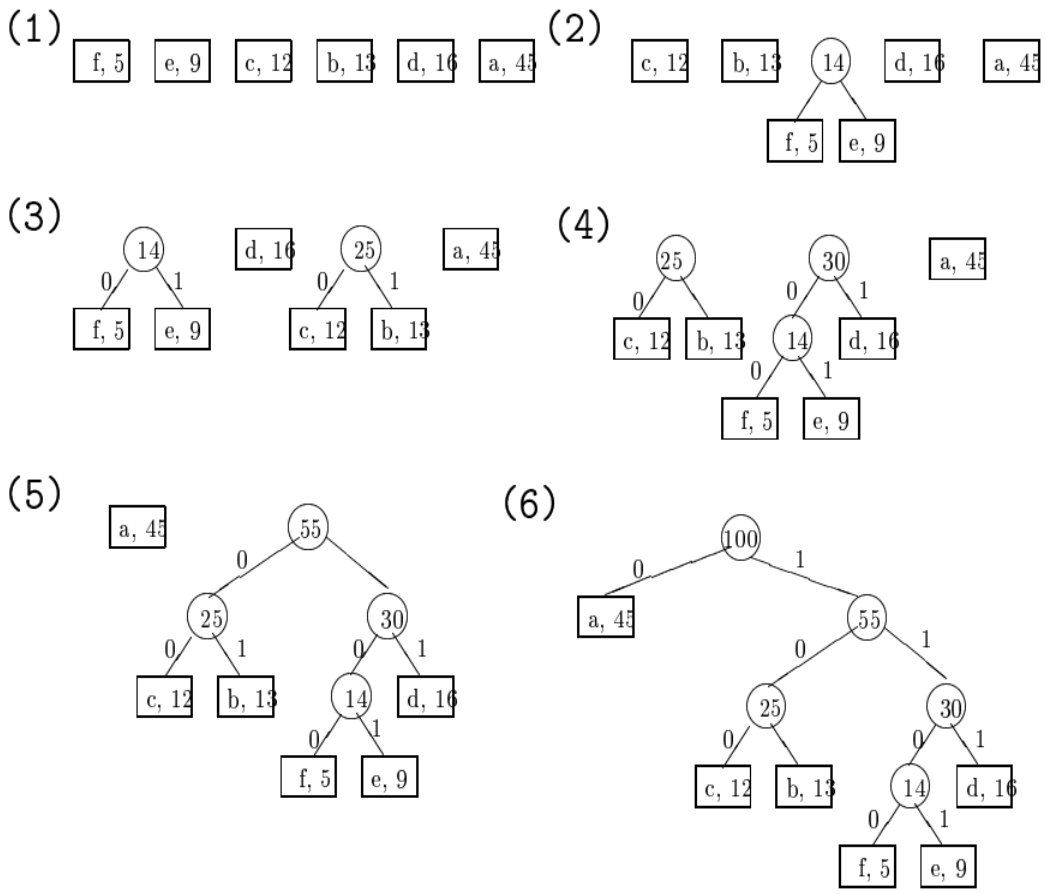
Αλγόριθμος Κωδικοποίησης Huffman

Ο πηγαίος κώδικας μεταβλητού μήκους που ελαχιστοποιεί το μέσο μήκος λήφθηκε από τον D. Huffman, ως αποτέλεσμα ενός προβλήματος εργασίας για το σπίτι που ανατέθηκε στην τάξη θεωρίας πληροφοριών του MIT από τον R. Fano. Η κωδικοποίηση Huffman είναι ένας αλγόριθμος συμπίεσης δεδομένων χωρίς απώλειες και αποτελεί έναν προθεματικό κώδικα με αναμενόμενο μήκος μικρότερο και από εκείνο του Shannon. Σε αυτόν τον αλγόριθμο, εκχωρείται ένας κωδικός μεταβλητού μήκους για την εισαγωγή διαφορετικών χαρακτήρων. Το μήκος του κώδικα σχετίζεται με το πόσο συχνά χρησιμοποιούνται χαρακτήρες. Οι πιο συχνοί χαρακτήρες έχουν τους μικρότερους κωδικούς και τους μεγαλύτερους κωδικούς για τους λιγότερο συχνούς χαρακτήρες.

Η πρακτικότητα του κώδικα Huffman έχει αντέξει στη δοκιμασία του χρόνου με μυριάδες εφαρμογές που κυμαίνονται από φαξ έως τηλεόραση υψηλής ευκρίνειας. Σύμφωνα με αυτό, για τη δυαδική κωδικοποίηση των συμβόλων της πηγής απαιτείται η εξής μεθοδολογία:

1. Πρώτα εντοπίζονται οι συχνότητες εμφάνισης ή πιθανότητες για κάθε σύμβολο.
2. Για κάθε ξεχωριστό σύμβολο δημιουργείται ένας κόμβος. Η συχνότητα αυτού του συμβόλου αποθηκεύεται μέσα στον κόμβο.
3. Εντοπίζονται οι δύο λιγότερο συχνοί κόμβοι (έστω x, y).
4. Διαμορφώνεται ένας νέος κοινός κόμβος z για τους δύο λιγότερο συχνούς κόμβους. Η συχνότητας (ή πιθανότητα) του συγκεκριμένου κόμβου είναι το άθροισμα των δύο. Ως αποτέλεσμα, έρχεται η μείωση κατά ένα του πλήθους των συμβόλων του αλφαβήτου της πηγής.
5. Επαναλαμβάνεται η διαδικασία έως ότου το αλφάβητο της πηγής αποτελείται μόνο από δύο σύμβολα, στα οποία αποδίδονται οι τιμές 0 και 1.
6. Ένα «0» και ένα «1» αποδίδονται στη θέση του ενός και του άλλου συμβόλου, αντίστοιχα, τα οποία στο βήμα 4 συγχωνεύτηκαν σε ένα. Το βήμα αυτό αφορά σε όλες τις συγχωνεύσεις.

Στο Διάγραμμα 3.1, δίνεται ένα οπτικό παράδειγμα των εν λόγο κόμβων (Nodes) του αλγορίθμου Huffman αποτελούμενοι από τα αρχικά σύμβολα πηγής και τις συχνότητες εμφάνισης (ή πιθανότητες) τους. Η διαδικασία ξεκινά από το βήμα (1), όπου αναπαριστά τα ξεχωριστά σύμβολα με τις αντίστοιχες μετρημένες συχνότητες εμφάνισης τους. Με την πάροδο των βημάτων του αλγόριθμου σχηματίζονται μοναδική κόμβοι αναπαράστασης των δύο σπανιότερων συμβόλων. Η διαδικασία αυτή επαναλαμβάνεται έως και το βήμα (6), με αποτέλεσμα να σχηματιστεί ένας και μοναδικός κόμβος.



Πηγή: Northwestern University (https://courses.cs.northwestern.edu/311/notes/huffman.pdf)

Διάγραμμα 3.1 - Παράδειγμα κωδικοποίησης Huffman

3.4 Διακριτές Πηγές Πληροφορίας Με Μνήμη

Όπως αναφέρθηκε στις προηγούμενές ενότητες οι διακριτές πηγές χωρίζονται σε εκείνες χωρίς μνήμη και με μνήμη. Η πρώτη κατηγορία, όπου και αναλύθηκε ήδη αφορά εκείνες τις πηγές που παράγουν ακολουθίες συμβόλων με τυχαίο και στατιστικά ανεξάρτητο τρόπο, ενώ στην άλλη πλευρά του νομίσματος βρίσκονται οι διακριτές πηγές που παράγουν σε διακριτές χρονικές στιγμές διαδοχικές ακολουθίες συμβόλων εξαρτημένες η μία από την άλλη. Η εξαρτημένη αυτή διαδικασία είναι φυσικά δυνατό να υφίσταται για μακρές ακολουθίες συμβόλων άλλα για πρακτικούς λόγους συνηθίζεται να εξετάζεται για περιορισμένο αριθμό συμβόλων. Αυτή η διαδικασία ανάλυσης της εξάρτησης είναι εφικτό να πραγματοποιηθεί μέσα από ειδικά στατιστικά υποδείγματα, γνωστά με την ονομασία Μαρκοβιανές Αλυσίδες *(Markov Chains)*. Στην συνέχεια, θα περιγράφουν λεπτομερώς οι έννοιες των μαρκοβιανών αλυσίδων και η συνεισφορά τους στην ανάλυση των πηγών πληροφορίας.

3.4.1 Μακροβιανές Αλυσίδες και Πηγές Markov

Οι Μακροβιανές Αλυσίδες καθίστανται στατιστικά υποδείγματα ικανά να ερμηνεύσουν διακριτές πηγές με μνήμη (εξάρτηση). Οι διακριτές πηγές οι οποίες έχουν την δυνατότητα να μοντελοποιηθούν σε Μαρκοβιανές Αλυσίδες ονομάζονται Πηγές Markov *(Markov Sources)*. Αναλυτικότερα, έστω μια ακολουθία τυχαίων μεταβλητών $\sum Y_i$ με $i = 1,2, \dots, n$ όπου υφίσταται κάποιου είδους εξάρτηση μεταξύ της παραγωγής αυτών των τυχαίων μεταβλητών της ακολουθίας. Η Μαρκοβιανή Αλυσίδα αποτελεί μια διαδικασία εξάρτησης μεταξύ τυχαίων μεταβλητών. Συγκεκριμένα, μια τυχαία μεταβλητή (έστω Y_i) εμφανίζει σημάδια εξάρτησης από την αμέσως προηγούμενη τυχαία μεταβλητή (έστω Y_{i-1}) στην ακολουθία ενώ ταυτόχρονα, είναι υπό συνθήκη ανεξάρτητη από όλες τις υπόλοιπες τυχαίες μεταβλητές. Επομένως, ισχύει η εξής σχέση:

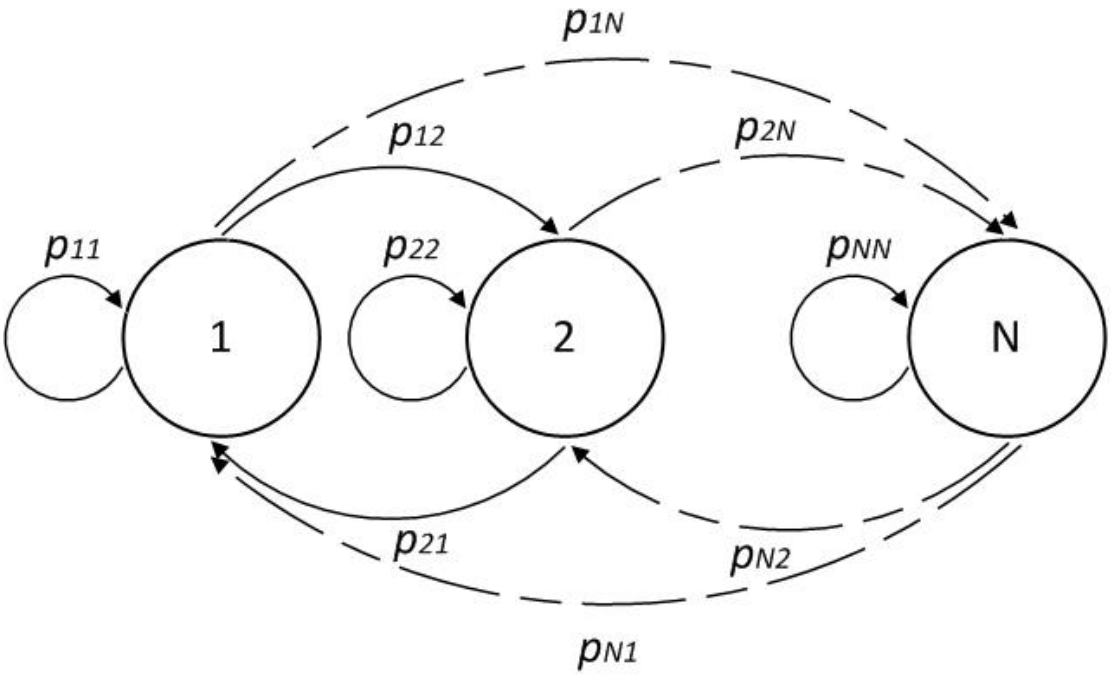
$$P(Y_{n+1} = y_{n+1} | Y_n = y_n, Y_{n-1} = y_{n-1}, \dots, Y_1 = y_1) = P(Y_{n+1} = y_{n+1} | Y_n = y_n)$$

Στη συγκεκριμένη διαδικασία, η συνάρτηση πιθανότητας μάζας θα γραφεί ως:

$$p(y_1, y_2, \dots, y_n) = p(y_1)p(y_2|y_1)p(y_3|y_2) \dots p(y_n|y_{n-1})$$

Όπου σχηματίζονται οι πιθανότητες μιας μεταβλητής y_i δεδομένου πως έχει παραχθεί η προηγούμενη της (y_{i-1}).

Αν και μια πηγή Markov εκπέμπει σύμβολα σε διάφορα χρονικά διαστήματα, γίνεται η υπόθεση της παραγωγής ενός συμβόλου σε ένα καθορισμένο χρονικό διάστημα, αναφερόμενο ως διάστημα συμβόλου. Σε κάθε χρονικό διάστημα συμβόλου η πηγή αλλάζει κατάσταση, δηλαδή μετατοπίζεται από μια αρχική κατάσταση του διαστήματος συμβόλου σε αυτή που συνεπάγεται το σύμβολο που εκπέμφθηκε. Η αρχική κατάσταση, η οποία αποτελεί την αρχή του διαστήματος, λαμβάνει δυνατές τιμές από m δυνατές καταστάσεις. Η πιθανότητα της μετάπτωσης από μια αρχική κατάσταση έστω i στην τελική έστω j, είναι ίση με P_{ij} και εξαρτάται μόνο από την αρχική και τη μελλοντική θέση του διαστήματος. Γενικά, η διαδικασία μετάπτωσης αναπαρίσταται με το διάγραμμα καταστάσεων της πηγής όπως διακρίνεται στο Διάγραμμα 3.2. Οι διάφορες δυνατές καταστάσεις ταυτίζονται με κάθε ένα κόμβο $(1,2,..n)$, ενώ οι ακμές $(i,j \in m)$ αναπαριστούν τις δυνατές μεταβιβάσεις καταστάσεων και με p_{ij} οι σχηματιζόμενες πιθανότητες.



Πηγή: Research Gate, 2018

Διάγραμμα 3.2 Διάγραμμα καταστάσεων πηγής τρίτης τάξης

Όπως γίνεται εύκολα αντιληπτό, εάν μία μαρκοβιανή αλυσίδα διακριτού χρόνου βρεθεί κάποια στιγμή στην κατάσταση i, τότε είναι αρκετά πιθανό την επόμενη στιγμή να βρεθεί στην κατάσταση j. Για παράδειγμα, μια αρχική κατάσταση είναι στο κόμβο 1, οι δυνατές επιλογές μετάπτωσης είναι από το 1 στο 2 και από το 1 στο κάθε N με πιθανότητες p_{12} ή p_{1N}.

Η διαδικασία Markov χαρακτηρίζεται ως **χρονικά** αμετάβλητη εάν η υπό συνθήκη πιθανότητα $p(y_{n+1}|y_n)$ δεν εξαρτάται από το αριθμό n. Δηλαδή:

$$P(Y_{n+1} = b|\, Y_n = a) \;=\; P(Y_2 = b|\, Y_1 = a)$$

Αν Yi είναι μια Μαρκοβιανή αλυσίδα, τότε η Yn αναπαριστά την κατάστασή της στο χρόνο n. Μια αμετάβλητη στο χρόνο Μαρκοβιανή αλυσίδα περιγράφεται πλήρως από την αρχική της κατάσταση και από τον πίνακα των πιθανοτήτων μετάβασης, δηλαδή $P = [Pij]$, όπου $Pij = P\{Y_{n+1} = j|Y_n = i\}$ και i, j οι τιμές των τυχαίων μεταβλητών οι οποίες ανήκουν στο σύνολο των δυνατών καταστάσεων $\{1, 2, \dots, m\}$. Ο πίνακας των πιθανοτήτων μετάβασης καλείται και πίνακας μετάπτωσης. Η πιθανότητα της Μαρκοβιανής αλυσίδας να βρίσκεται τη χρονική στιγμή n στην κατάσταση i συμβολίζεται με $p_i(n)$, δηλαδή $p_i(n) = P_i(Y_n = i)$. Αν ισχύει $p_i(n) = p_i(n + 1) = \pi_i$ για κάθε κατάσταση, τότε η Μαρκοβιανή αλυσίδα χαρακτηρίζεται στατική. Για μια στατική Μαρκοβιανή αλυσίδα ισχύει μεταξύ του διανύσματος των πιθανοτήτων των καταστάσεων π και του πίνακα μετάπτωσης P η σχέση $\pi P \;=\; \pi$.

Η διακριτή τυχαία μεταβλητή Yi αναπαριστά την κατάσταση της διαδικασίας για τα σύμβολα που εκπέμφθηκαν στα τελευταία l διαστήματα συμβόλων, όπου l είναι ο αριθμός των προηγούμενων συμβόλων που επηρεάζουν το επόμενο που θα παραχθεί. Ουσιαστικά, η διαδικασία Markov αποτελεί ένα σύστημα μετατοπίσεων από μια κατάσταση σε μια νέα, όπου κάθε μια μετακίνηση επηρεάζει τόσο την επόμενη μετάπτωση όσο και ολόκληρο το σύστημα. Βάθος Πηγής είναι ακριβώς αυτός ο αριθμός ύπαρξης προηγούμενων συμβόλων επιρροής. Το πλήθος των δυνατών καταστάσεων στην πηγή υπολογίζεται μέσω του πλήθους των συμβόλων της ίδιας της πηγής (q) και του βάθους πηγής, δηλαδή:

$$m = q^l$$

Στην υπόθεση πως το βάθος πηγής είναι $l = 1$, άρα η μαρκοβιανή αλυσίδα χαρακτηρίζεται πρώτης τάξης, τότε το πλήθος των δυνατών καταστάσεων της πηγής είναι ίσο με το πλήθος των συμβόλων του αλφαβήτου της πηγής:

$$m = q^1$$

Επεξηγηματικά, εάν για πρώτης τάξης μαρκοβιανης αλυσίδας, η πηγής βρίσκεται στην κατάσταση i με μετάπτωση στην j, ισχύει ότι έχει λάβει χώρα ως τελευταία εκπομπή αυτή του συμβόλου s_i με μελλοντική εκπομπή του s_j. Για βάθος πηγής $l = 2$, όπου η επιλογή του υπό εκπομπή συμβόλου εξαρτάται μόνο από τα δύο τελευταία σύμβολα, το πλήθος των δυνατών καταστάσεων της πηγής θα είναι $m = q^2$.

Για τις χρονικά αμετάβλητες διαδικασίες Markov, οι πιθανότητες των καταστάσεων της πηγής υπολογίζονται βάση της παρακάτω εξίσωσης:

$$p_j(k+1) = \sum_{i=1}^{m} p_i(k) P_{ij}$$

όπου $p_i(k)$ είναι η πιθανότητα να βρίσκεται το σύστημα στην κατάσταση i κατά την αρχή του διαστήματος συμβόλου k.

3.4.2 Εντροπία των Πηγών Markov

Η μέση ποσότητα πληροφορία ή εντροπία μιας πηγής Markov είναι ο μέσος όρος της εντροπίας των συμβόλων που εκπέμπονται από κάθε κατάσταση. Για μια κατάσταση i, η εντροπία των συμβόλων εκτιμάται από την εξής σχέση:

$$H(K_i) = -\sum_{j=1}^{m} P_{ij} \, log \, P_{ij} \;\; bits/symbol$$

Επομένως, η εντροπία της πηγής σχηματίζεται από το μέσο όρο της εντροπίας των καταστάσεων, δηλαδή:

$$H(S) = \sum_{i=1}^{m} p_i H(K_i) = -\sum_{i=1}^{m} p_i \sum_{j=1}^{m} P_{ij} \, log \, P_{ij}$$

Από την παραπάνω σχέση, ορίζεται ο μέσος ρυθμός πληροφορίας της πηγής R με r_s, το ρυθμό εκπομπής συμβόλων από την πηγή. Άρα:

$$R = r_s H(s)\ bits/symbol$$

Η μέση ποσότητα πληροφορίας μηνυμάτων της πηγής, δηλαδή το άθροισμα όλων των μηνυμάτων μήκους l συμβόλων και $p(m_i)$, η πιθανότητα εκπομπής του μηνύματος m_i, ορίζεται από την σχέση:

$$H(M) = -\sum p(m_i)\ log\ p(m_i) \qquad 3.1$$

Από την Εξίσωση 3.1 γίνεται εύκολα ο προσδιορισμός της μέσης ποσότητας πληροφορίας συμβόλων της πηγής εάν διαιρεθεί με το μήκος των μηνυμάτων ως εξής:

$$H_l = \frac{1}{l} H(M)$$

Σημειώνεται, πως κατά ανάλογο τρόπο με τις διακριτές πηγές χωρίς μνήμη εξετάζεται και στις διακριτές πήγες με μνήμη η απομάκρυνση του πλεοναστικού περιεχομένου ενός μηνύματος κατά την κωδικοποίηση των πηγών Markov. Συγκεκριμένα, ο πλεονασμός στις διακριτές πηγές με μνήμη διακρίνεται σε δύο κατηγορίες, εκείνος της εξάρτησης και ο ολικός. Το μέτρο πλεονασμού εξάρτησης διαμορφώνεται μέσω της σχέσης:

$$red_{\varepsilon\xi} = 1 - \frac{H_{\mu\varepsilon\ \mu\nu\acute{\eta}\mu\eta}(S)}{H_{\chi\omega\rho\acute{\iota}\varsigma\ \mu\nu\acute{\eta}\mu\eta}(S)}$$

Από την άλλη, το μέτρο ολικού πλεονασμού αναφέρεται στην εντροπία της πηγής με μνήμη συγκριτικά με τη μέγιστη δυνατή εντροπία της πηγής χωρίς μνήμη που επιτυγχάνεται για ίσες πιθανότητες εκπομπής όλων των συμβόλων. Άρα:

$$red_{o\lambda} = 1 - \frac{H_{\mu\varepsilon\ \mu\nu\acute{\eta}\mu\eta}(S)}{max H_{\chi\omega\rho\acute{\iota}\varsigma\ \mu\nu\acute{\eta}\mu\eta}(S)} = 1 - \frac{H_{\mu\varepsilon\ \mu\nu\acute{\eta}\mu\eta}(S)}{log\ q}$$

Με αυτόν τον τρόπο, γίνεται εύκολη η διαδικασία απομάκρυνσης του πλεονασμού κατά την κωδικοποίηση μηνύματος στην πηγή, διαμορφώνοντας ένα ουσιαστικό και ορθό μήνυμα ικανό να αναγνωστεί στα επόμενα στάδια ενός ψηφιακού επικοινωνιακού δικτύου.

3.5 Ανακεφαλαίωση

Το κεφάλαιο αυτό επικεντρώθηκε στην περιγραφή των δύο κατηγοριών των διακριτών πηγών καθώς και τις τεχνικές κωδικοποίησης των πηγών αυτών για βέλτιστη συμπίεση της πληροφορίας. Συγκεκριμένα, αναλύθηκαν οι διακριτές πηγές χωρίς μνήμη, ο εντοπισμός της εντροπίας, του ρυθμού μετάδοσης ενός μηνύματος, καθώς και ο εντοπισμός του πλεοναστικού περιεχομένου αυτού. Κατόπιν αναλύθηκε η διαδικασία κωδικοποίησης ή συμπίεσης μιας πληροφορίας με σκοπό την απομάκρυνση του μη απαραίτητου περιεχομένου ενός μηνύματος για ταχύτερη και αποδοτικότερη μετάδοση. Τέλος, αντίστοιχη ανάλυση έγινε για τις φημισμένες Μαρκοβιανές Πηγες και την μοντελοποίηση των διακριτών πηγών με μνήμη σε μαρκοβιανα υποδείγματα ανάλυσης.

ΚΕΦΑΛΑΙΟ 4

ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ

4.1 Εισαγωγή

Η κωδικοποίηση πηγής αποτελεί αναπόσπαστο μέρος της ψηφιακής επικοινωνίας για μετάδοση μιας πληροφορίας από τον πομπό (υπολογιστή Α) προς τον δέκτη (υπολογιστή Β). Όπως έχει αναφερθεί ήδη στα προηγούμενα κεφάλαια, δεν είναι εφικτή η απευθείας μετάδοση της γνώσης με το αρχικό αλφάβητο χρήσης $(\alpha, \beta, \gamma, \delta)$ αλλά απαιτείται μετατροπή σε αλφάβητο αναγνωρίσιμο από το ψηφιακό μηχάνημα $(0, 1)$. Η κωδικοποίηση χαρακτήρων είναι η διαδικασία αντιπροσώπευσης μεμονωμένων χαρακτήρων ενός αλφάβητου χρησιμοποιώντας ένα αντίστοιχο σύστημα κωδικοποίησης που αποτελείται από διαφορετικά σύμβολα (δυαδικό αλφάβητο) Μέσω αυτής της διαδικασίας επιτυγχάνεται η απομάκρυνση πλεοναστικού περιεχομένου του εκάστοτε μηνύματος για διευκόλυνση αποθήκευσης, διαχείρισης και μεταφοράς κειμένου μέσω τηλεπικοινωνιακών δικτύων.

4.2 Εκτέλεση Lembel Ziv Αλγορίθμου Κωδικοποίησης

Στην συνέχεια εκτελείται ένα παράδειγμα κωδικοποίησης αλγορίθμου στην πηγή για συμπίεση δεδομένων. Ο αλγόριθμος Lempel-Ziv είναι ένας αλγόριθμος συμπίεσης δεδομένων, που αναπτύχθηκε από τους Abraham Lempel και Jacob Ziv στη δεκαετία του 1970. Αναλυτικά, ο αλγόριθμος αποτελείται από μια οικογένεια αλγορίθμων, συγκεκριμένα των LZ77 και LZ78, με σκοπό τη συμπίεση και την αποσυμπίεση αρχικών δεδομένων. Συγκεκριμένα, εκμεταλλεύονται την επαναληπτική δομή των δεδομένων για να τα συμπιέσουν . Οι αλγόριθμοι Lempel-Ziv είναι γενικοί και μπορούν να χρησιμοποιηθούν για τη συμπίεση κειμένων, εικόνων, ήχου και άλλων μορφών δεδομένων. Θεωρείται ένας από τους πιο γνωστούς και ευρέως χρησιμοποιούμενους αλγορίθμους συμπίεσης, και έχει εφαρμοστεί σε πολλά δημοφιλή συστήματα συμπίεσης δεδομένων, όπως το αρχείο ZIP και οι εικόνες GIF. Παρακάτω δίνεται ο αλγόριθμος του Lempel-Ziv που έτρεξα στο MATLAB R2022α.

```
clc
clear

alphabets = [0 1]; % Διακριτά σύμβολα που μπορεί να παράγει η πηγή δεδομένων
p = [.97 .03]; % Πιθανοτική κατανομή για τα σύμβολα
n = 100 * 1024; % Μήκος της εισαγωγικής ακολουθίας

x = zeros(1, n);
cumProb = cumsum(p);

% Δημιουργία της εισαγωγικής ακολουθίας βάση της πιθανοτικής κατανομής
for i = 1:n
    randNum = rand;
    x(i) = alphabets(find(randNum <= cumProb, 1));
end

disp('================================');
disp('Lempel-Ziv input sequence is created');
strInput = strrep((mat2str(x)), ' ', '');
strInput = strrep(strInput, '[', '');
strInput = strrep(strInput, ']', '');
disp('mat2str completed');

codeBook = cellstr(['0'; '1']); % Αρχικοποίηση του codebook με τα αρχικά σύμβολα

% Εφαρμογή του αλγορίθμου Lempel – Ziv στην εισαγωγική ακολουθία
[value, codeBook, NumRep, NumRepBin] = lempelzivEnc(strInput, codeBook);

outputLength = length(NumRepBin{1}) * length(NumRepBin) - length(NumRepBin); %
Υπολογισμός του μήκους της συμπιεσμένης εξόδου
inputLength = length(strInput); % Υπολογισμός του μήκους της εισαγωγικής ακολουθίας
compRatio = outputLength / inputLength * 100; % Υπολογισμός του ποσοστού συμπίεσης
str = sprintf('======================\nInput length is %d and output length is
%d\nCompression ratio is %f', inputLength, outputLength, compRatio);
disp(str);

% Συνάρτηση κωδικοποίησης του Lempel-Ziv
function [output, outCodeBook, NumRep, NumRepBin] = lempelzivEnc(Inputdata,
codeBook)
    disp('Start of Lempel-Ziv encoder');
    buffer = '';
    output = '';
    searchIndex = 0;
    codeBookLength = [];
    for i = 1:length(codeBook)
        codeBookLength(i) = length(codeBook{i});
    end
    str = sprintf('Total input length is %d', length(Inputdata));
    disp(str);
    prevPos = -1;
    for i = 1:length(Inputdata)
        c = Inputdata(i);
        searchIndex = 0;
        if mod(i, 10240) == 0
            time1 = clock;
            str = sprintf('%s - Completed length is %d KBit', datestr(now), i /
1024);
```

```
            disp(str);
        end

        codeWord = [buffer c];
        codeLength = length(codeWord);

        % Αναζήτηση της λέξης στο codebook
        for j = 1:length(codeBook)
            if codeBookLength(j) == codeLength
                if codeBook{j} == (codeWord)
                    searchIndex = j;
                    break;
                end
            end
        end

        if searchIndex ~= 0
            % Η κωδική λέξη υπάρχει ήδη στο codebook
            buffer = codeWord;
            prevPos = searchIndex;
        else
            % Η κωδική λέξη δεν υπάρχει στο codebook
            startIndex = length(codeBook) + 1;
            codeBook{startIndex} = codeWord;
            codeBookLength(startIndex) = length(codeWord);
            output = [output num2str(prevPos) ','];
            buffer = '';
        end
    end

    outCodeBook = codeBook;

    disp('Lempel-Ziv encoder is completed.');
    str = sprintf('Total length of code book is %d', length(codeBook));
    disp(str);

    NumRep = [];
    NumRepBin = [];
    wordLength = ceil(log2(length(codeBook) - 2));

    % Δημιουργία αριθμητικής και δυαδικής αναπαράστασης των καταχωρήσεων του
codebook
    for i = 3:length(codeBook)
        if mod(i, 3000) == 0
            str = sprintf('Current CodeBook binary representation is %d', i);
            disp(str);
        end

        strTofind = '';
        strToFindRight = '';
        if codeBookLength(i) == 1
            strTofind = codeBook{i};
        else
            strTofind = codeBook{i}(1:codeBookLength(i) - 1);
            strToFindRight = codeBook{i}(codeBookLength(i):codeBookLength(i));
        end
```

```
        firstPosition = 0;
        secondPosition = 0;

        % Εύρεση των δεικτών των καταχωρήσεων του codebook για την αριθμητική και
δυαδική αναπαράσταση
        for j = 1:i
            if length(strTofind) == codeBookLength(j)
                if strTofind == codeBook{j}
                    firstPosition = j;
                end
            end
            if length(strToFindRight) == codeBookLength(j)
                if strToFindRight == codeBook{j}
                    secondPosition = j;
                end
            end
            if firstPosition ~= 0 && secondPosition ~= 0
                break;
            end
        end

        NumRep{length(NumRep) + 1} = [num2str(firstPosition) ','
num2str(secondPosition)];
        NumRepBin{length(NumRepBin) + 1} = [num2str(dec2bin(firstPosition,
wordLength)) ',' codeBook{secondPosition}];
    end
end
```

Παρακάτω παρουσιάζονται αναλυτικά τα βήματα εκτέλεσης του αλγορίθμου Lempel-Ziv.

Αρχικοποίηση:

1) Το αλφάβητο των συμβόλων που μπορεί να παράγει η πηγή δεδομένων ορίζεται ως [0, 1].
2) Η κατανομή πιθανότητας των συμβόλων καθορίζεται ως [0,97, 0,03], όπου το 0,97 αντιπροσωπεύει την πιθανότητα του συμβόλου 0 και το 0,03 αντιπροσωπεύει την πιθανότητα του συμβόλου 1.
3) Το μήκος της ακολουθίας εισόδου ορίζεται σε n = 100 * 1024.

Δημιουργία της ακολουθίας εισόδου:

Η ακολουθία εισόδου δημιουργείται με δειγματοληψία συμβόλων από την καθορισμένη κατανομή πιθανότητας. Το μήκος της ακολουθίας εισόδου είναι n και κάθε σύμβολο επιλέγεται με βάση την αντίστοιχη πιθανότητα.

Κωδικοποίηση Lempel-Ziv:

1) Η διαδικασία κωδικοποίησης Lempel-Ziv ξεκινά με την προετοιμασία ενός βιβλίου κωδικών με τα αρχικά σύμβολα [0, 1].
2) Η ακολουθία εισόδου διασχίζεται σύμβολο με σύμβολο.
3) Για κάθε σύμβολο που συναντάται, σχηματίζεται μια κωδική λέξη με τη σύνδεση του τρέχοντος συμβόλου με ένα buffer που περιέχει τα προηγούμενα σύμβολα.
4) Στη συνέχεια, η κωδική λέξη συγκρίνεται με τις εγγραφές στο βιβλίο κωδικών για να προσδιοριστεί εάν υπάρχει ήδη.
5) Εάν η κωδική λέξη βρεθεί στο βιβλίο κωδικών, το buffer ενημερώνεται με την τρέχουσα κωδική λέξη και το ευρετήριο της κωδικής λέξης στο βιβλίο κωδικών καταγράφεται ως η προηγούμενη θέση.
6) Εάν η κωδική λέξη δεν βρεθεί στο βιβλίο κωδικών, η κωδική λέξη προστίθεται στο βιβλίο κωδικών, καταγράφεται το ευρετήριο της προηγούμενης θέσης και η προσωρινή μνήμη διαγράφεται.
7) Η διαδικασία συνεχίζεται έως ότου υποβληθούν σε επεξεργασία όλα τα σύμβολα στην ακολουθία εισόδου.
8) Η έξοδος της διαδικασίας κωδικοποίησης Lempel-Ziv είναι μια σειρά από δείκτες που αντιπροσωπεύουν τις θέσεις των λέξεων κώδικα στο βιβλίο κωδικών, σχηματίζοντας μια συμπιεσμένη αναπαράσταση της ακολουθίας εισόδου.

Υπολογισμός του μήκους εξόδου και του λόγου συμπίεσης:

1) Το μήκος της συμπιεσμένης εξόδου υπολογίζεται με βάση τον αριθμό των δεικτών κωδικών λέξεων.
2) Το μήκος της ακολουθίας εισόδου και της ακολουθίας εξόδου χρησιμοποιούνται για τον υπολογισμό του λόγου συμπίεσης.

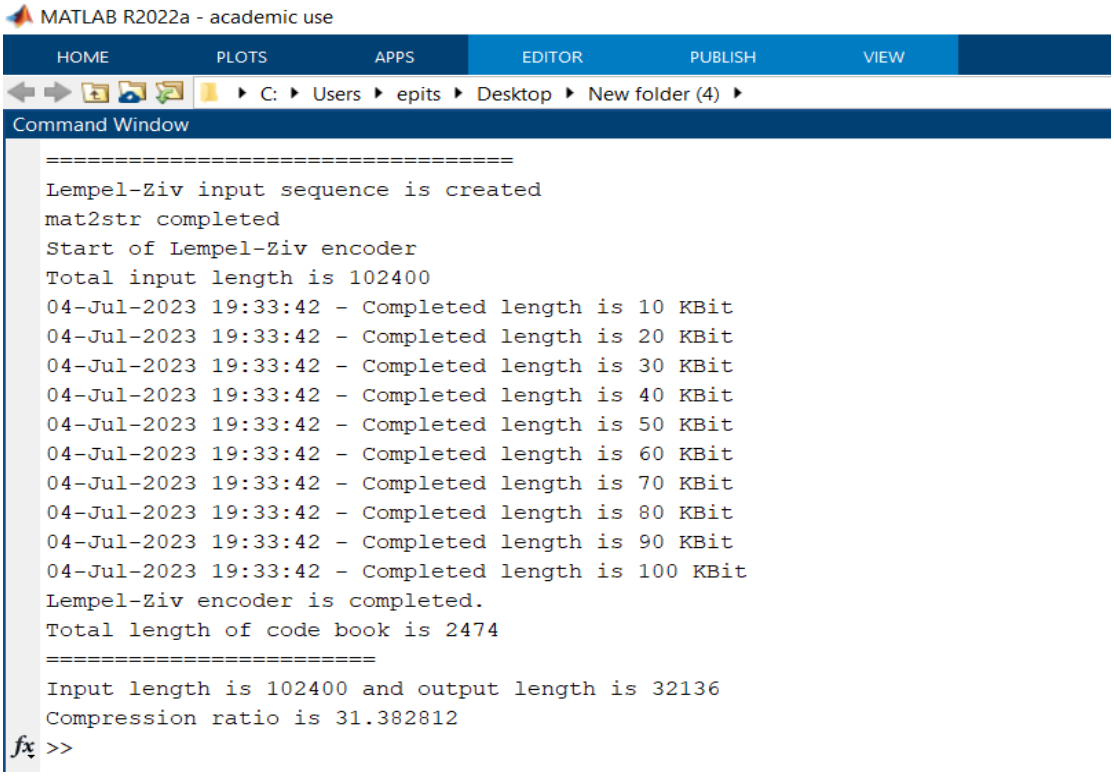
Λειτουργία κωδικοποίησης Lempel-Ziv:

1) Η διαδικασία κωδικοποίησης Lempel-Ziv είναι ενσωματωμένη στη συνάρτηση lempelzivEnc.
2) Η συνάρτηση λαμβάνει ως είσοδο την ακολουθία εισόδου και το βιβλίο κωδικών.
3) Αρχικοποιεί μεταβλητές και δομές δεδομένων για την αποθήκευση του buffer, της εξόδου, του βιβλίου κωδικών και άλλων πληροφοριών.
4) Η συνάρτηση επαναλαμβάνεται πάνω από την ακολουθία εισόδου και εκτελεί τα βήματα κωδικοποίησης που περιγράφηκαν προηγουμένως.

5) Δημιουργεί επίσης αριθμητικές και δυαδικές αναπαραστάσεις των εγγραφών του βιβλίου κωδικών, οι οποίες αποθηκεύονται για περαιτέρω ανάλυση.

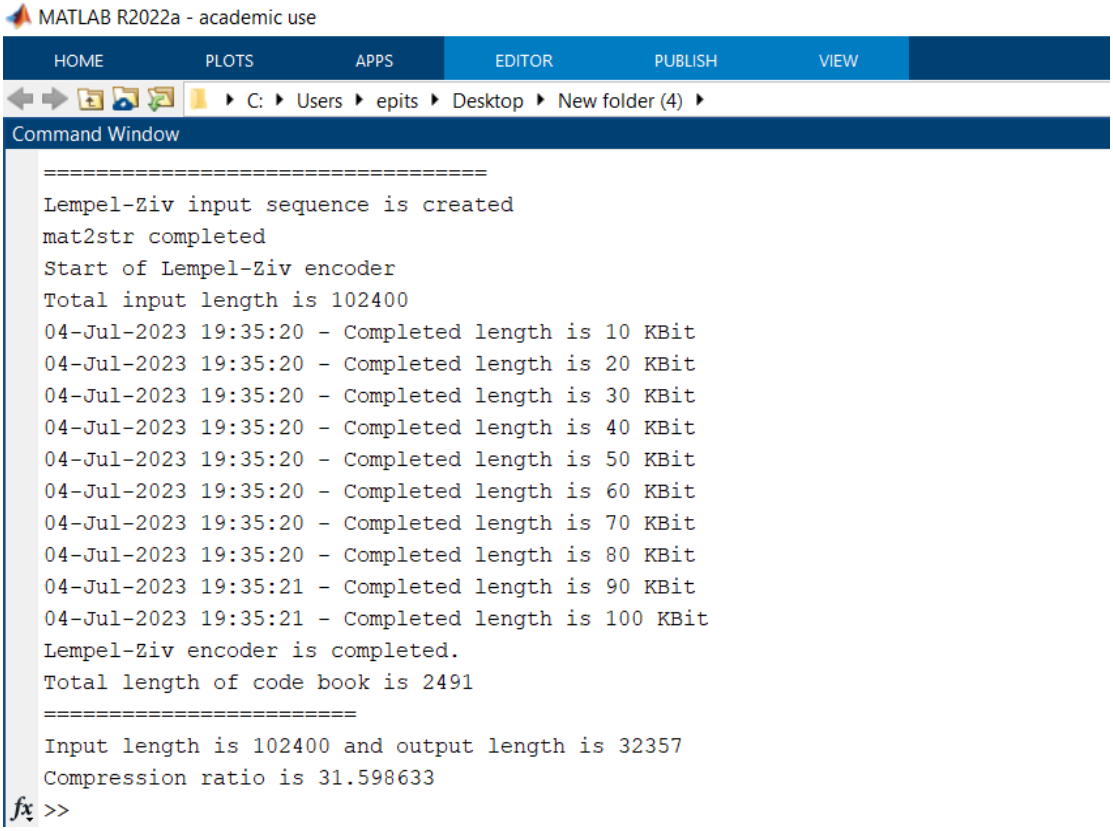
Κατόπιν, εκτελέστηκε ο αλγόριθμος δύο φορές και στην συνέχεια παρουσιάζονται τα εξής αποτελέσματα:

Πίνακας 4.1 Αποτελέσματα Πρώτης Εκτέλεσης Αλγορίθμου Lempel-Ziv



Σε αυτό το παράδειγμα έχουμε 2474 κωδικές λέξεις που αποθηκεύονται στον πίνακα CodeBook. Το input length μας δείχνει τον αριθμό των συμβόλων που παράχθηκαν από την πηγή όπου στον αλγόριθμο αυτόν είναι σταθερά 100*1024 δηλαδή 102.400 σύμβολα (bits). Έπειτα έχουμε το output length το οποίο είναι το μήκος της κωδικοποιημένης ακολουθίας και είναι 32136 σύμβολα (bits). Τέλος για να βρούμε το ποσοστό συμπίεσης της ακολουθίας διαιρούμε το output length με το input length και το πολλαπλασιάζουμε με το 100 και παίρνουμε το ποσοστό δηλαδή το 31,382812%.

Πίνακας 4.2 Αποτελέσματα Δεύτερης Εκτέλεσης Αλγορίθμου Lempel-Ziv



Σε αυτό το παράδειγμα έχουμε 2491 κωδικές λέξεις που αποθηκεύονται στον πίνακα CodeBook. Το input length μας δείχνει τον αριθμό των συμβόλων που παράχθηκαν από την πηγή όπου στον αλγόριθμο αυτόν είναι σταθερά 100*1024 δηλαδή 102.400 σύμβολα (bits). Έπειτα έχουμε το output length το οποίο είναι το μήκος της κωδικοποιημένης ακολουθίας και είναι 32357 σύμβολα (bits). Τέλος για να βρούμε το ποσοστό συμπίεσης της ακολουθίας διαιρούμε το output length με το input length και το πολλαπλασιάζουμε με το 100 και παίρνουμε το ποσοστό δηλαδή το 31,598633%. Όπως φαίνεται από την δεύτερη εκτέλεση του αλγορίθμου έχουμε μικρότερο ποσοστό συμπίεσης.

Επίλογος

Στην παρούσα πτυχιακή εργασία περιγράφηκε ο κλάδος της Θεωρίας της Πληροφορίας. Στο πρώτο κεφάλαιο έγινε η ιστορική αναδρομή της θεωρίας της πληροφορίας αναφέροντας τους συμβαλλόμενους επιστήμονες. Αναλύθηκε η έννοια της πληροφορίας σε τρείς διαφορετικές κατηγορίες, κάθε μια από τις οποίες έχουν διαφορετικό νόημα και στόχο ενώ παράλληλα, αναφέρθηκε περιληπτικά η διαδικασία μετάδοσης της πληροφορίας, ο ορισμός των διακριτών πηγών επικοινωνίας μέσα στα οποία διαμορφώνεται ένα μήνυμα καθώς και η έννοια της κωδικοποίησης στο στάδιο της πηγής.

Στη συνέχεια, το δεύτερο κεφάλαιο περιέγραψε την θεωρία πιθανοτήτων και όλων των σχετιζόμενων εννοιών που την αφορούν. Συγκεκριμένα, ορίστηκε το θεώρημα του Bayes, διευκρινίστηκε το μέτρο πληροφορίας που όρισε ο Hartley το 1928 πάνω στο οποίο βασίστηκε η θεωρία του Shannon σχετικά με την μέση ποσότητα πληροφορίας.

Τέλος, το τελευταίο κεφάλαιο επικεντρώθηκε στην περιγραφή των δύο κατηγοριών των διακριτών πηγών καθώς και τις τεχνικές κωδικοποίησης των πηγών αυτών για βέλτιστη συμπίεση της πληροφορίας.

Η Θεωρία της Πληροφορίας, η οποία έχει βάσεις από την θεωρία πιθανοτήτων, χρησιμοποιείται αισθητά στον κλάδο των τηλεπικοινωνιών και έχει καίριες εφαρμογές στην ασφάλεια δικτύων. Ως εκ τούτου, η Θεωρία Πληροφορίας η οποία αναπτύχθηκε από τους εξαιρετικούς επιστήμονες Hartley και Shannon με τις εμπεριστατωμένες εργασίες τους πάνω στην έννοια, της ποσοτικοποίησης και την διάκριση της πληροφορίας, αποτελεί σήμερα τον επιστημονικό πυρήνα εξέτασης πολλών διαφορετικών επιστημών συμβάλλοντας έντονα στην περεταίρω τεκμηρίωση και ανάπτυξή τους.

Βιβλιογραφία

Manikas, A. (χ.χ.). Information Sources. Imperial College London.

Reza, F. M. (1994). *An Introduction to Information Theory.* Dover .

Roman, S. (1997). *Introduction to Coding and Information Theory.* Springer.

Thomas, T. M.-J. (2014). *Στοιχεία της θεωρίας πληροφορίας.* Πανεπιστημιακές Εκδόσεις Κρήτης.

University, N. (1989). Introdution to Algorithms - Huffman Coding. Northwestern University.

ΖΟΡΚΑΔΗΣ, Β. (2002). *Θεωρία Πληροφορίας και Κωδικοποίησης.* Πάτρα: ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ.

Μυλωνάς, Ν. (2013). *Πιθανότητες και στατιστική.* Τζιόλα.

https://hith aldia.ac.in

https://www.researchgate.net