ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ

Διαχείριση Πληροφορίας και Κατηγοριοποίηση Διεπαφών Αναζήτησης στον Παγκόσμιο Ιστό με Αλγόριθμους Εμπνευσμένους από τη Φύση και Τεχνικές Μηχανικής Μάθησης για Μεγάλα Δεδομένα

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Βασίλειος Γ. Κόλιας

Συμβουλευτική επιτροπή: Ελευθέριος Καγιάφας (Επιβλέπων)
Βασίλειος Λούμος
Ιωάννης Αναγνωστόπουλος

ΑΘΗΝΑ, 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ

Διαχείριση Πληροφορίας και Κατηγοριοποίηση Διεπαφών Αναζήτησης στον Παγκόσμιο Ιστό με Αλγόριθμους Εμπνευσμένους από τη Φύση και Τεχνικές Μηχανικής Μάθησης για Μεγάλα Δεδομένα

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Βασίλειος Γ. Κόλιας

Συμβουλευτική επιτροπή: Ελευθέριος Καγιάφας (Επιβλέπων)
Βασίλειος Λούμος
Ιωάννης Αναγνωστόπουλος

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή στις 7 Μαΐου 2015

.....	.....	.....
Ε. Καγιάφας	Β. Λούμος	Ι. Αναγνωστόπουλος
Καθηγητής	Καθηγητής	Επίκουρος Καθηγητής
Ε.Μ.Π.	Ε.Μ.Π.	Πανεπιστήμιο Θεσσαλίας
.....	.....	.....
Θ. Βαρβαρίγου	Α. Σταφυλοπάτης	Μ. Θεολόγου
Καθηγήτρια	Καθηγητής	Καθηγητής
Ε.Μ.Π.	Ε.Μ.Π.	Ε.Μ.Π.
	.....	
	Δημήτριος Βέργαδος	
	Επίκουρος Καθηγητής	
	Πανεπιστήμιο Πειραιώς	

.....

Βασίλειος Γ. Κόλιας

Περίληψη

Ένα μεγάλο ποσοστό του περιεχομένου στον Παγκόσμιο Ιστό είναι μη διαθέσιμο στους χρήστες των παραδοσιακών μηχανών αναζήτησης εξαιτίας της αδυναμίας προσπέλασής του από τα προγράμματα προσκομιδής περιεχομένου τους. Το φαινόμενο αυτό οφείλεται στο γεγονός ότι το περιεχόμενο αυτό παράγεται δυναμικά και μόνο σαν αποτέλεσμα της υποβολής ερωτημάτων σε φόρμες αναζήτησης. Η αναγνώριση των διεπαφών αυτών αποτελεί το πρώτο βήμα για την αυτοματοποιημένη προσπέλαση περιεχομένου σε αυτό το κομμάτι του Παγκόσμιου Ιστού που είναι γνωστό ως Κρυμμένος Παγκόσμιος Ιστός. Σε αυτή τη διατριβή, αρχικά παρουσιάζεται το εν λόγω ερευνητικό πεδίο. Στη συνέχεια γίνεται ανάλυση ενός συνόλου ιστοσελίδων μεγάλης κλίμακας, με στόχο την εξαγωγή χρήσιμων συμπερασμάτων για τις διεπαφές που περιέχονται στο σύνολο αυτό. Ταυτόχρονα, με βάση αυτό το σύνολο ιστοσελίδων κατασκευάστηκε ένα σύνολο εκπαίδευσης για την επαγωγή κανόνων κατηγοριοποίησης για την αυτοματοποιημένη αναγνώριση διεπαφών αναζήτησης.

Η επαγωγή κανόνων κατηγοριοποίησης είναι μια από τις παλιότερες τεχνικές μηχανικής μάθησης και έχει εφαρμοστεί επιτυχώς σε πολλαπλά προβλήματα. Το κύριο πλεονέκτημά της είναι η απλότητα του παραγόμενου μοντέλου κατηγοριοποίησης και η ευκολία ανάγνωσης και ερμηνείας του από τον ανθρώπινο παράγοντα. Μια από τις συνεισφορές της διατριβής αυτής είναι μια πρωτότυπη κατανεμημένη τεχνική επαγωγής κανόνων κατηγοριοποίησης βασισμένη στο μοντέλο Απεικόνισης/Μείωσης. Σαν πρώτο βήμα η προσέγγιση μετατρέπει τα δεδομένα εκπαίδευσης από συνεχή σε διακριτά και στη συνέχεια αναζητά εξαντλητικά το χώρο των πιθανών κανόνων για την εύρεση του καλύτερου, βασισμένη σε ένα προκαθορισμένο κριτήριο αξιολόγησης. Οι κανόνες που παράγονται από το παραπάνω σύνολο, χρησιμοποιούνται για την κατηγοριοποίηση διεπαφών αναζήτησης στον Παγκόσμιο Ιστό ως προς τη λειτουργία τους.

Τέλος παρουσιάζεται ένας πρωτότυπος αλγόριθμος εμπνευσμένος από φυσικές διεργασίες για την αναζήτηση πληροφορίας στον Παγκόσμιο Ιστό. Ο αλγόριθμος αυτός έχει τη δυνατότητα να εντοπίζει συναφείς πληροφοριακές μονάδες δρομολογώντας την αναζήτηση πληροφορίας μέσα στο δυναμικό περιβάλλον του Παγκόσμιου Ιστού. Η δρομολόγηση της αναζήτησης, πραγματοποιείται στοχαστικά συνδυάζοντας τεχνικές ανάκτησης που βασίζονται στην ομοιότητα εγγράφων και τεχνικών προσομοίωσης του τρόπου επικοινωνίας των μυρμηγκιών. Ο προτεινόμενος αλγόριθμος σε συνδυασμό με τους κανόνες κατηγοριοποίησης που παράγονται από την προηγούμενη προσέγγιση, μπορεί να εντοπίσει θεματικά συναφείς διεπαφές αναζήτησης στον Παγκόσμιο Ιστό για τη διευκόλυνση της αναζήτησης πληροφορίας στον Κρυμμένο Ιστό.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Κρυμμένος Παγκόσμιος Ιστός, Επαγωγή Κανόνων, Απεικόνιση/Μείωση, Αναζήτηση στον Παγκόσμιο Ιστό, Αποικία Μυρμηγκιών

Abstract

A large portion of the content residing on the Web, though publicly available, is inaccessible by the traditional general purpose search engines, since it is only generated dynamically as a response to a query submission to a search interface. The identification of such interfaces is the first step towards the automated access to the content of the part of the Web known as Hidden Web. In this dissertation, after an introduction to this research field, a large scale analysis is conducted on a publicly available dataset containing web pages with complex interfaces. The aim of the analysis is to draw useful conclusions on the properties of the interfaces, and the construction of a training dataset for the induction of classification rules that will be used for the automated identification of search interfaces.

Classification rule induction is one of the oldest machine learning techniques and it has been successfully applied in a plethora of problems. Its main advantage is the simplicity of the resulting classification model and the ability for a human to interpret it. One of the contributions of this dissertation, is the introduction of a novel distributed classification rule induction approach based on MapReduce. As a first step, the approach transforms any numeric attributes to discrete and then it exhaustively searches the space of possible rules to find the best one according to an evaluation criterion. The resulting rules are used for the functional classification of interfaces.

Finally a novel nature inspired algorithm for searching for information on the Web is also presented. The proposed algorithm has the ability to locate relative information units by routing the search in the dynamic environment of the Web. The search is conducted stochastically, by combining techniques based on document similarity and techniques that emulate the communication of real world ants in their foraging process. The proposed algorithm when used in conjunction with the classification rules induced previously can locate, similar search interfaces on the Web in order to facilitate the access to Hidden Web content.

KEYWORDS:

Hidden Web, Rule Induction, MapReduce, Web Search, Ant Colony

Ευχαριστίες

Η διδακτορική διατριβή αυτή περιλαμβάνει το σύνολο της ερευνητικής μου δραστηριότητας, η οποία εκπονήθηκε στο Εργαστήριο Τεχνολογίας Πολυμέσων του Εθνικού Μετσόβιου Πολυτεχνείου. Η τελική της μορφή οφείλεται και στη βοήθεια και την υποστήριξη συνεργατών, προς τους οποίους κρίνω σκόπιμο να απευθύνω τις ευχαριστίες μου.

Καταρχάς θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες προς τους επιβλέποντές μου Καθηγητές Ελευθέριο Καγιάφα και Βασίλειο Λούμο. Η ευκαιρία που μου έδωσαν για την εκπόνηση του διδακτορικού μου στο Εργαστήριο Τεχνολογίας Πολυμέσων, η καθοδήγηση, οι συμβουλές τους αλλά και η εμπιστοσύνη που έδειξαν στις δυνατότητές μου κατά το χρονικό διάστημα της συνεργασίας μας, ήταν καθοριστικές. Ιδιαίτερα, θα ήθελα να ευχαριστήσω τον Επίκουρο Καθηγητή Ιωάννη Αναγνωστόπουλο για την καθοδήγηση που μου προσέφερε και την εμπιστοσύνη που μου έδειξε σε όλα τα στάδια αυτής της προσπάθειας αλλά και για την ουσιαστική συνεισφορά του στη διατριβή. Επίσης, θα ήθελα να εκφράσω τις ευχαριστίες μου στο Δρ. Γεώργιο Κούζα για την εισαγωγή στο θεματικό πεδίο της διατριβής που μου παρείχε, αλλά και την καθοριστική συμβολή του στην εκπόνησή της.

Θα ήθελα να ευχαριστήσω τους Καθηγητές Θ. Βαρβαρίγου, Α. Σταφυλοπάτη, Μ. Θεολόγου και τον Επίκουρο Καθηγητή Δ. Βέργαδο για την τιμή που μου προσέφεραν συμμετέχοντας στην επταμελή επιτροπή κρίσης της διατριβής μου.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου για τη στήριξή τους όλα αυτά τα χρόνια και τον αδερφό μου Κωνσταντίνο, για την αδιάκοπη συμπαράσταση που μου παρείχε σε όλη τη σταδιοδρομία μου.

Βασίλειος Γ. Κόλιας,

Μάιος 2015

Περιεχόμενα

Ευρετήριο Εικόνων

Ευρετήριο Πινάκων

Ευρετήριο Αλγορίθμων

Εισαγωγή

Αντικείμενο της διατριβής

Βασικό αντικείμενο της διατριβής αυτής είναι η χρήση μιας πρωτότυπης κατανεμημένης τεχνικής επαγωγής κανόνων κατηγοριοποίησης βασισμένης στο μοντέλο Απεικόνισης/Μείωσης για την κατηγοριοποίηση διεπαφών αναζήτησης στον Παγκόσμιο Ιστό ως προς το είδος της λειτουργίας τους. Ένα μεγάλο ποσοστό του περιεχομένου του Παγκόσμιου Ιστού είναι απροσπέλαστο από τις παραδοσιακές μηχανές αναζήτησης αφού δημιουργείται δυναμικά σαν αποτέλεσμα της υποβολής ερωτημάτων σε τέτοιες διεπαφές. Η αναγνώριση διεπαφών αναζήτησης αποτελεί το πρώτο βήμα για την αυτοματοποιημένη προσπέλαση περιεχομένου σε αυτό το κομμάτι του Παγκόσμιου Ιστού που καλείται Κρυμμένος Παγκόσμιος Ιστός. Μετά την παρουσίαση του εν λόγω ερευνητικού πεδίου, γίνεται μια ανάλυση ενός συνόλου ιστοσελίδων σε μεγάλη κλίμακα με στόχο την εξαγωγή χρήσιμων συμπερασμάτων για τις διεπαφές αλλά και την κατασκευή του συνόλου εκπαίδευσης της προτεινόμενης προσέγγισης. Μια ακόμα συνεισφορά της διατριβής αυτής είναι η παρουσίαση ενός πρωτότυπου αλγόριθμου εμπνευσμένου από φυσικές διεργασίες για την αναζήτηση πληροφορίας στον Παγκόσμιο Ιστό. Ο αλγόριθμος αυτός σε συνδυασμό με τους κανόνες κατηγοριοποίησης που παράγονται από την προηγούμενη προσέγγιση, μπορεί να εντοπίσει θεματικά συναφείς διεπαφές αναζήτησης στον Παγκόσμιο Ιστό για τη διευκόλυνση της αναζήτησης πληροφορίας σε αυτόν.

Περιγραφή του προβλήματος

Στις σύγχρονες πληροφοριακές κοινωνίες, ο Παγκόσμιος Ιστός αποτελεί αναπόσπαστο κομμάτι των καθημερινών δραστηριοτήτων εκατομμύριων ανθρώπων. Τόσο ο τεράστιος όγκος δεδομένων που υπάρχει στον Παγκόσμιο Ιστό όσο και η ποσότητα των δεδομένων που μπορεί να υπάρξει σε ένα μεμονωμένο ιστότοπο, έχει καταστήσει αδύνατη την εύρεση πληροφοριών με χειροκίνητο τρόπο. Η μόνη φιλική προς το χρήστη μέθοδος για την ικανοποίηση των πληροφοριακών αναγκών του, είναι η χρήση αυτοματοποιημένων μηχανών αναζήτησης. Κάθε μηχανή αναζήτησης στον Ιστό διαθέτει τουλάχιστον μια φόρμα αναζήτησης, μέσω της οποίας επιτρέπεται η υποβολή συγκεκριμένων ερωτημάτων, βάσει των οποίων σχηματίζονται οι σελίδες με τα αντίστοιχα αποτελέσματα. Ωστόσο, η δυναμική αυτή παραγωγή περιεχομένου σαν αποτέλεσμα της υποβολής ερωτημάτων μέσω μιας φόρμας αναζήτησης, αποτελεί σημαντικό εμπόδιο στη διαδικασία προσκομιδής περιεχομένου των παραδοσιακών μηχανών αναζήτησης γενικού σκοπού. Αυτό έχει σαν αποτέλεσμα να μην εντάσσεται στα ευρετήρια των μηχανών αναζήτησης ένας σημαντικός όγκος περιεχομένου και κατ' επέκταση να μην συμπεριλαμβάνεται σαν πιθανό αποτέλεσμα κατά την υποβολή ερωτημάτων από τους χρήστες.

Κάθε μέθοδος επίλυσης του συγκεκριμένου προβλήματος που έχει προταθεί στη βιβλιογραφία μέχρι σήμερα, περιλαμβάνει σαν πρώτο βήμα την αναγνώριση των διεπαφών που προορίζονται για αναζήτηση έναντι αυτών που προορίζονται για οποιοδήποτε άλλο σκοπό, όπως την εγγραφή, τη σύνδεση ή τη δημοσίευση ενός σχολίου. Ο μεγαλύτερος όγκος της βιβλιογραφίας για την αναγνώριση των διεπαφών, στηρίζεται σε ευρετικούς κανόνες που έχουν παραχθεί χειροκίνητα βάσει της εμπειρίας ειδικών στο χώρο. Σε αυτή τη διατριβή, το πρόβλημα της αναγνώρισης διεπαφών αναζήτησης αντιμετωπίζεται σαν ένα πρόβλημα κατηγοριοποίησης από τη σκοπιά μιας μηχανής αναζήτησης γενικού σκοπού. Σε αυτή την περίπτωση, πρέπει να αναλυθεί μαζικά ένας τεράστιος όγκος ιστοσελίδων που έχουν συλλεχθεί από προγράμματα προσκομιδής περιεχομένου γενικού σκοπού σε κάποιο κεντρικό αποθετήριο. Κατ' αναλογία του σεναρίου αυτού, στη διατριβή αυτή αναλύεται το σύνολο δεδομένων Yahoo L11 από την υπηρεσία Webscope της Yahoo που περιέχει ένα σύνολο ιστοσελίδων με διεπαφές. Το σύνολο αυτό είναι δημόσια διαθέσιμο μέσω της υπηρεσίας S3 της Amazon. Η ανάλυση του Yahoo L11 γίνεται με τεχνικές επεξεργασίας Μεγάλων Δεδομένων και έχει σαν στόχο την εξαγωγή συμπερασμάτων σχετικά με τα χαρακτηριστικά των διεπαφών αναζήτησης και τις ιδιότητες που μπορούν να χρησιμοποιηθούν για την εκπαίδευση αλγόριθμων μηχανικής μάθησης. Κύρια μέριμνα κατά τον προσδιορισμό των ιδιοτήτων αυτών είναι η ανεξαρτησία από τη φυσική γλώσσα συγγραφής των ιστοσελίδων και των διεπαφών, έτσι ώστε ο εκπαιδευμένος κατηγοριοποιητής να μπορεί να χρησιμοποιηθεί σε όλον τον Παγκόσμιο Ιστό, χωρίς να χρειάζεται επανεκπαίδευση για κάθε φυσική γλώσσα.

Στη διατριβή αυτή επίσης προτείνεται μια προσέγγιση επαγωγής κανόνων βασισμένη στην τεχνική Απεικόνισης/Μείωσης. Η επαγωγή κανόνων κατηγοριοποίησης είναι μια από τις παλιότερες τεχνικές μηχανικής μάθησης και έχει εφαρμοστεί επιτυχώς σε πολλαπλά προβλήματα. Το κύριο πλεονέκτημά της είναι η απλότητα του παραγόμενου μοντέλου κατηγοριοποίησης και η ευκολία ανάγνωσης και ερμηνείας του από τον ανθρώπινο παράγοντα. Ωστόσο παρά την εξέλιξη και ωριμότητα του πεδίου της επαγωγής κανόνων, υπάρχουν ελάχιστες προσεγγίσεις που έχουν υλοποιηθεί οι οποίες λαμβάνουν υπόψη το μέγεθος του συνόλου εκπαίδευσης, έτσι ώστε να κλιμακώνονται με την αύξησή του. Προς αυτό τον στόχο, στην προτεινόμενη προσέγγιση χρησιμοποιείται το μοντέλο της Απεικόνισης/Μείωσης που αποτελεί μια πολύ διαδεδομένη προσέγγιση για την επεξεργασία δεδομένων μεγάλης κλίμακας. Παρά τους περιορισμούς του συγκεκριμένου μοντέλου, η προτεινόμενη προσέγγιση, μέσα από μια σειρά πειραματικών αξιολογήσεων, αποδεικνύεται ότι είναι αποδοτική και κλιμακωτή και ότι μπορεί κατ' επέκταση να εφαρμοστεί σε άλλα προβλήματα. Η συγκεκριμένη προσέγγιση εφαρμόζεται στο σύνολο δεδομένων του Yahoo L11 και παράγεται ένα σύνολο κανόνων οι οποίοι μέσα από μια σειρά πειραμάτων, δείχνεται ότι έχουν μεγάλη ακρίβεια στην κατηγοριοποίηση διεπαφών. Κατά συνέπεια, μπορούν να χρησιμοποιηθούν για την αναγνώριση διεπαφών αναζήτησης, στα πλαίσια μιας προσέγγισης με στόχο την προσκομιδή περιεχομένου από τον Κρυμμένο Ιστό.

Τέλος, στη διατριβή αυτή, προτείνεται μια προσέγγιση στο πρόβλημα της αναζήτησης πληροφορίας στον Παγκόσμιο Ιστό που βασίζεται στον αλγόριθμο αποικίας μυρμηγκιών. Η προσέγγιση ενεργεί με την υπόθεση ότι σχετικές ιστοσελίδες βρίσκονται σε κοντινή απόσταση, αν ως μέτρο απόστασης θεωρηθεί το πλήθος υπερσυνδέσμων που χρειάζεται να ακολουθηθούν από μια ιστοσελίδα σε μια άλλη. Η προτεινόμενη προσέγγιση αναζητά σχετικές ιστοσελίδες γύρω από μια αρχική πηγή, παρόμοια με τη διεργασία αναζήτησης τροφής των μυρμηγκιών γύρω από μια αποικία. Η τεχνική αναζήτησης βασίζεται στον αλγόριθμο αποικίας μυρμηγκιών που έχει εφαρμοστεί επιτυχώς σε μια πλειάδα προβλημάτων όπως αυτά της βελτιστοποίησης ή της επαγωγής κανόνων κατηγοριοποίησης και η σύγκριση της ομοιότητας δυο ιστοσελίδων γίνεται βάσει του χωροδιανυσματικού μοντέλου. Απώτερος και μελλοντικός στόχος της δουλειάς αυτής είναι να συνδυαστεί η συγκεκριμένη προσέγγιση αναζήτησης με έναν εκπαιδευμένο κατηγοριοποιητή έτσι ώστε να οδηγήσει στην προσκομιδή διεπαφών αναζήτησης συναφών εννοιολογικά.

Δομή της διατριβής

Στο πρώτο κεφάλαιο της διατριβής γίνεται μια εισαγωγή στην έννοια των Μεγάλων Δεδομένων. Αφού δοθεί ο ορισμός των Μεγάλων Δεδομένων και η περιγραφή των κύριων χαρακτηριστικών τους, γίνεται μια εκτενής παρουσίαση των μεθόδων επεξεργασίας τους με ιδιαίτερη έμφαση στο μοντέλο Απεικόνισης/Μείωσης που αποτελεί τη βάση πάνω στην οποία υλοποιείται η προσέγγιση επαγωγής κανόνων που περιγράφεται στο δεύτερο κεφάλαιο και τη μέθοδο επεξεργασίας του συνόλου δεδομένων Yahoo L11 που παρουσιάζεται στο τρίτο κεφάλαιο της διατριβής.

Στο δεύτερο κεφάλαιο περιγράφεται η υλοποίηση ενός αλγόριθμου επαγωγής κανόνων για κατηγοριοποίηση δεδομένων που βασίζεται στο μοντέλο Απεικόνισης/Μείωσης, με στόχο την τελική κατασκευή ενός αποδοτικού μοντέλου κατηγοριοποίησης σε όσο το δυνατόν λιγότερες επαναλήψεις. Επίσης περιγράφεται η υλοποίηση μιας προσέγγισης διακριτοποίησης ως ένα βήμα προεπεξεργασίας των δεδομένων το οποίο καθιστά δυνατή την επαγωγή κανόνων από αυτά. Μετά από μια αναλυτική περιγραφή των δυο προσεγγίσεων, η προτεινόμενη λύση αξιολογείται από τρεις οπτικές γωνίες: α) την ακρίβειά της ως προς την κατηγοριοποίηση, β) την επίδοσή της ως προς την παράλληλη εκτέλεση και γ) το κόστος στην επικοινωνία μεταξύ των επεξεργαστικών μονάδων. Η αξιολόγηση υποδεικνύει ότι η προσέγγιση είναι κλιμακούμενη και επειδή παράγει μοντέλα εύκολα κατανοητά στους ανθρώπους, μπορεί να αποδειχθεί χρήσιμη σε μια πλειάδα εφαρμογών.

Στο τρίτο κεφάλαιο, μετά από τον ορισμό και την περιγραφή των κυριότερων χαρακτηριστικών του Κρυμμένου Ιστού, γίνεται μια επισκόπηση των προσεγγίσεων που έχουν προταθεί στη βιβλιογραφία για την προσπέλαση του περιεχομένου του. Το πρώτο βήμα σε κάθε μια από αυτές τις προσεγγίσεις είναι η αναγνώριση των διεπαφών ως προς τη λειτουργία τους. Οι προσεγγίσεις αυτές χρειάζονται σαν είσοδο, διεπαφές που όντως προορίζονται για αναζήτηση και όχι για άλλο σκοπό,

όπως την εγγραφή ή τη σύνδεση σε έναν ιστότοπο, την υποβολή κάποιου σχολίου κλπ. Η αναγνώριση των διεπαφών αναζήτησης τους αντιμετωπίζεται σαν πρόβλημα κατηγοριοποίησης και για την επίλυσή του χρησιμοποιείται ο αλγόριθμος επαγωγής κανόνων που περιγράφτηκε στο Κεφάλαιο 2. Για την εκπαίδευση του κατηγοριοποιητή, επιλέχθηκε ένα δείγμα από το σύνολο των διεπαφών που εξήχθησαν από το σύνολο δεδομένων Yahoo L11. Μετά από ανάλυση του συγκεκριμένου συνόλου, επιλέγονται οι ιδιότητες του συνόλου εκπαίδευσης, με κύριο στόχο να είναι ανεξάρτητες της φυσικής γλώσσας στην οποία είναι γραμμένες, έτσι ώστε το παραγόμενο μοντέλο να μπορεί να εφαρμοστεί σε όλον τον Παγκόσμιο Ιστό. Τέλος μετά από μια εμπειρική επιλογή των πιο κατάλληλων παραμέτρων για την προσέγγιση, γίνεται μια πειραματική μελέτη για την αξιολόγηση της επίδοσής της και μια σύγκριση με άλλους αλγόριθμους κατηγοριοποίησης.

Τέλος στο τέταρτο κεφάλαιο προτείνεται μια νέα μεθοδολογία αναζήτησης πληροφορίας στον Παγκόσμιο Ιστό που στηρίζεται σε μια από τις πιο γνωστές τεχνικές εμπνευσμένες από τη φύση, της αποικίας μυρμηγκιών. Πιο συγκεκριμένα, προτείνεται μια προσέγγιση με το όνομα Ant-Seeker που έχει τη δυνατότητα να εντοπίζει συναφείς πληροφοριακές μονάδες δρομολογώντας την αναζήτηση πληροφορίας μέσα στο δυναμικό περιβάλλον του Παγκόσμιου Ιστού. Η δρομολόγηση της αναζήτησης, πραγματοποιείται στοχαστικά συνδυάζοντας τεχνικές ανάκτησης που βασίζονται στην ομοιότητα εγγράφων και τεχνικών προσομοίωσης του τρόπου επικοινωνίας των μυρμηγκιών. Αφού παρουσιαστούν αρχικά οι βασικότερες εμπνευσμένες από τη φύση τεχνικές, γίνεται ανάλυση των περιορισμών των παραδοσιακών μεθόδων μηχανικής μάθησης καθώς και των λόγων που οδήγησαν στη αναζήτηση τέτοιων εναλλακτικών, μη συμβατικών μεθόδων. Παράλληλα παρουσιάζονται οι αρχές δύο θεμελιωδών μεθόδων εμπνευσμένων από τη φύση. Στη συνέχεια παρουσιάζεται ο προτεινόμενος αλγόριθμος και το κεφάλαιο κλείνει με την παρουσίαση πειραματικών μετρήσεων για την προτεινόμενη μεθοδολογία καθώς και ποιοτικά συμπεράσματα.

Κεφάλαιο 1

Μεγάλα Δεδομένα και τεχνικές επεξεργασίας αυτών

1.1 Εισαγωγή

Στο πρώτο κεφάλαιο της διατριβής αυτής γίνεται μια εισαγωγή στην έννοια των Μεγάλων Δεδομένων. Αφού δοθεί ο ορισμός των Μεγάλων Δεδομένων και η περιγραφή των κύριων χαρακτηριστικών τους, γίνεται μια εκτενής παρουσίαση των μεθόδων επεξεργασίας τους με ιδιαίτερη έμφαση στο μοντέλο Απεικόνισης/Μείωσης που αποτελεί τη βάση πάνω στην οποία υλοποιείται η προσέγγιση επαγωγής κανόνων που περιγράφεται στο δεύτερο κεφάλαιο και η μέθοδος ανάλυσης του συνόλου δεδομένων Yahoo L11 που παρουσιάζεται στο τρίτο κεφάλαιο της διατριβής.

Η αξία της ανάλυσης Μεγάλων Δεδομένων έγκειται αφενός στη βελτίωση της στατιστικής αξιοπιστίας καθώς αυξάνεται ο όγκος των δεδομένων και αφετέρου στη βελτίωση των στατιστικών μοντέλων καθαυτών, αφού καθίσταται δυνατός ο εντοπισμός συσχετισμένων παραγόντων, που θα ήταν εξαιρετικά δύσκολο να γίνει με δεδομένα σε μικρότερες κλίμακες. Με άλλα λόγια τα Μεγάλα Δεδομένα παρέχουν μια πιο καθολική άποψη των δεδομένων ενός προβλήματος. Η ανάλυση Μεγάλων Δεδομένων ωστόσο θα πρέπει να γίνεται προσεκτικά, καθώς μαζί με την ανάδειξη συσχετιζόμενων παραγόντων, μπορεί να αναδείξει συσχετιζόμενους θορύβους. Κάτι τέτοιο μπορεί να αποφευχθεί με τη χρήση κατάλληλων μεθόδων και εργαλείων για τη συλλογή, την αποθήκευση, τη μετάδοση και τη γραφική αναπαράσταση των δεδομένων προς ανάλυση.

1.2 Ορισμός των Μεγάλων Δεδομένων

Στη σύγχρονη εποχή, οι ερευνητικές διεργασίες στα περισσότερα πεδία εφαρμογών, καθοδηγούνται και υποστηρίζονται από δεδομένα. Η αξία των δεδομένων καθορίζεται αφενός από την αφθονία τους αφετέρου από τη χρήσιμη πληροφορία που μπορεί να εξαχθεί από αυτά. Οι πρόσφατες εξελίξεις στις τεχνολογίες μετάδοσης δεδομένων και την κινητή υπολογιστική, διευκόλυναν τόσο την παραγωγή των δεδομένων όσο και τη συλλογή τους. Ταυτόχρονα, οι εξελίξεις στις κατανεμημένες επεξεργασίες και η εμφάνιση της υπολογιστικής νέφους (cloud computing), βελτίωσαν τόσο τις δυνατότητες αποθήκευσης των δεδομένων όσο και τις δυνατότητες επεξεργασίας τους. Αυτό είχε σαν αποτέλεσμα την ανάδειξη των τεράστιων δυνατοτήτων που προκύπτουν από την ανάλυση συνόλων δεδομένων μεγάλης κλίμακας, όχι επιλεκτικά σε κάποιο δείγμα τους, αλλά στην ολότητά τους. Οι παράγοντες αυτοί, είχαν σαν αποτέλεσμα το σχηματισμό ενός νέου ερευνητικού πεδίου, αυτού των Μεγάλων Δεδομένων (Big Data), που πλέον βρίσκεται στο επίκεντρο του

ενδιαφέροντος τόσο της ερευνητικής κοινότητας της Επιστήμης των Υπολογιστών, όσο και της βιομηχανίας αλλά και του ευρύτερου κοινού γενικότερα.

Ο όρος Μεγάλα Δεδομένα προέκυψε από τρεις ξεχωριστές περιοχές: την έρευνα, τη βιομηχανία και τα μέσα ενημέρωσης και για το λόγο αυτό δεν υπάρχει ένας κοινά αποδεκτός ορισμός τους. Τα Μεγάλα Δεδομένα έχουν κατά κύριο λόγο συνδεθεί με δύο ιδέες: α) την αποθήκευση δεδομένων και β) τη συστηματική ανάλυσή τους. Οι δύο προαναφερθείσες ιδέες ωστόσο δεν είναι κάτι το νέο, αφού εδώ και πολλές δεκαετίες έχουν αποτελέσει αντικείμενο ερευνών που έχουν οδηγήσει σε αμέτρητες εφαρμογές. Η ουσιώδης διαφορά μεταξύ των Μεγάλων Δεδομένων και των συμβατικών μεθόδων επεξεργασίας δεδομένων που έχουν αναπτυχθεί μέχρι τώρα, εστιάζεται στον όρο "Μεγάλα". Το επίθετο "Μεγάλα" γενικότερα είναι ποσοτικό προσδιοριστικό, ωστόσο στα πλαίσια της έννοιας των "Μεγάλων Δεδομένων" αναφέρεται στη σπουδαιότητα, την πολυπλοκότητα και τις προκλήσεις που εγείρονται σε όσες διαδικασίες έχουν να αντιμετωπίσουν, με τον έναν ή τον άλλο τρόπο, δεδομένα. Εύλογα λοιπόν προκαλείται σύγχυση και δυσκολία στην απόδοση του ορισμού τους. Σύμφωνα με την εργασία [1] έχουν δοθεί αρκετοί ορισμοί από διαφορετικές πηγές οι επικρατέστεροι από τους οποίους είναι οι ακόλουθοι:

Από τους πιο συχνά παρατεθειμένους ορισμούς για τα Μεγάλα Δεδομένα είναι αυτός του ερευνητικού οργανισμού Gartner [2], όπου παρόλο που δεν γίνεται ρητή αναφορά στον όρο, παρέχεται ένα σύνολο χαρακτηριστικών που θα πρέπει τα δεδομένα αυτά να φέρουν για να μπορούν να χαρακτηριστούν ως τέτοια. Τα χαρακτηριστικά αυτά είναι γνωστά ως "τα τρία V": α) όγκος (volume), β) ταχύτητα (velocity) και γ) η ποικιλία (variety). Ο ορισμός αυτός δίνει έμφαση στο αυξανόμενο μέγεθος των δεδομένων, τον αυξανόμενο ρυθμό παραγωγής τους και το αυξανόμενο εύρος της μορφής και της δομής των δεδομένων που αναλύονται. Τα τρία αυτά χαρακτηριστικά παρουσιάζονται εμπειρικά, χωρίς εμπεριστατωμένες μετρήσεις από πραγματικές εφαρμογές, αφού άλλωστε κάτι τέτοιο δεν είναι δυνατόν. Ο ορισμός αυτός έχει αναθεωρηθεί από το NIST [3] και τον οργανισμό Gartner [4], ενώ έχει επεκταθεί από την IBM [5] και άλλους με το να συμπεριλαμβάνει και ένα τέταρτο V: την ακρίβεια στην απόδοση της αλήθειας (veracity). Το τελευταίο αυτό χαρακτηριστικό περιλαμβάνει τα θέματα της αξιοπιστίας και της εμπιστοσύνης που προκύπτουν από την αβεβαιότητα που συνοδεύει τα δεδομένα και τα αποτελέσματα της ανάλυσής τους.

Ο ορισμός που παρέχεται από την Oracle [6], αποφεύγει τη χρήση των παραπάνω χαρακτηριστικών και επικεντρώνεται στην αξία που προκύπτει από τη διαδικασία λήψης αποφάσεων που βασίζεται σε δομημένα δεδομένα και που ενισχύεται από νέες πηγές αδόμητων δεδομένων. Τέτοιες πηγές περιλαμβάνουν ιστολόγια, κοινωνικά μέσα, δίκτυα αισθητήρων, δεδομένα εικόνων και άλλες μορφές δεδομένων που ποικίλλουν ως προς το μέγεθός τους, τη δομή τους και άλλα χαρακτηριστικά. Κατά συνέπεια ο ορισμός αυτός δίνει έμφαση στην επέκταση των πηγών δεδομένων ως μέσο υποβοήθησης των υπαρχουσών διαδικασιών. Επίσης ο ορισμός της Oracle, επικεντρώνεται στις υποδομές που αποτελούν λύση στα θέματα που προκύπτουν από τα Μεγάλα Δεδομένα, όπως

ένα σύνολο τεχνολογιών που περιλαμβάνουν βάσεις δεδομένων NoSQL, το οικοσύστημα του Hadoop, το πακέτο R και τις σχεσιακές βάσεις δεδομένων.

Ο ορισμός που παρέχεται από την Intel [7], είναι ο μοναδικός που συνδέει τα Μεγάλα Δεδομένα με απτές ποσότητες. Η Intel προσδιορίζει ως Μεγάλα Δεδομένα αυτά που προκύπτουν από οργανισμούς που παράγουν σαν μέση ποσότητα δεδομένων 300TB την εβδομάδα. Η περιγραφή αυτή προκύπτει ως ποσοτικοποίηση των εμπειριών της εταιρείας με τους εταιρικούς συνεργάτες της. Όπως και η Oracle, η Intel αποκαλύπτει ότι ο πιο συχνός τύπος δεδομένων που χρησιμοποιείται κατά την επεξεργασία Μεγάλων Δεδομένων είναι σχεσιακά δεδομένα, ενώ ακολουθούν έγγραφα, email, δεδομένα αισθητήρων, ιστολόγια και κοινωνικά μέσα.

Τέλος η Microsoft [8] παρέχει έναν περιληπτικό ορισμό: Ο όρος Μεγάλα Δεδομένα χρησιμοποιείται όλο και περισσότερο για την περιγραφή της χρήσης ολοένα και περισσότερης υπολογιστικής ισχύος κυρίως για τους σκοπούς της μηχανικής μάθησης και της τεχνητής νοημοσύνης σε σύνολα δεδομένων μεγάλης κλίμακας και μεγάλης πολυπλοκότητας. Αυτός ο ορισμός ξεκαθαρίζει ότι η επεξεργασία των Μεγάλων Δεδομένων χρειάζεται σημαντική υπολογιστική ισχύ, κάτι που δεν κατέστη σαφές από τους προηγούμενους ορισμούς. Επιπρόσθετα ο ορισμός δίνει έμφαση σε δυο τεχνολογίες: την μηχανική μάθηση και την τεχνητή νοημοσύνη που δεν αναφέρονται από κανέναν άλλο ορισμό.

Πέραν των διαφορών που υπάρχουν στους διάφορους ορισμούς που έχουν διατυπωθεί στη βιβλιογραφία, υπάρχουν κάποια σημεία στα οποία συγκλίνουν:

- Το μέγεθος του συνόλου δεδομένων προς ανάλυση είναι ενδεχομένως ο σημαντικότερος παράγοντας
- Η πολυπλοκότητα, η δομή και η συμπεριφορά των δεδομένων είναι υψηλής σημασίας
- Τα εργαλεία και οι τεχνικές που χρησιμοποιούνται για την επεξεργασία των δεδομένων παίζουν σημαντικό ρόλο

Απόρροια της ποικιλίας των τύπων δεδομένων που υπάρχουν είναι η ποικιλομορφία των πληροφοριών που μπορούν να εξαχθούν από αυτά. Το γεγονός αυτό δημιουργεί την ανάγκη ύπαρξης εξειδικευμένων τεχνικών για την εξαγωγή τους. Οι τεχνικές αυτές εντάσσονται στον ευρύτερο όρο των τεχνικών ανάλυσης δεδομένων (data analytics) που περιλαμβάνει τεχνικές για την αντιμετώπιση των προκλήσεων που προκύπτουν από την αναζήτηση, την ανάλυση και την απεικόνιση των δεδομένων που έχουν σαν απώτερο στόχο την εξαγωγή χρήσιμων πληροφοριών για την καθοδήγηση της λήψης αποφάσεων. Προαπαιτούμενο των τεχνικών ανάλυσης ωστόσο, είναι η είσοδος των δεδομένων στις διαδικασίες αυτές και το ρόλο αυτό παίζουν τα συστήματα υποδομής που περιλαμβάνουν τεχνικές λήψης, αποθήκευσης και μετάδοσης δεδομένων.

1.3 Μοντέλα επεξεργασίας Μεγάλων Δεδομένων

Τα συστήματα επεξεργασίας και διαχείρισης Μεγάλων Δεδομένων είναι κατασκευασμένα κατά τέτοιο τρόπο ώστε να εκμεταλλεύονται τα παράλληλα ή κατανεμημένα συστήματα για να ανταποκρίνονται αποδοτικά σε κλιμακωτές απαιτήσεις. Συνήθως σε τέτοια συστήματα, ο φόρτος εργασίας κατανέμεται σε πολλαπλούς υπολογιστικούς κόμβους, αναθέτοντας διαφορετικά κομμάτια των δεδομένων σε διαφορετικούς επεξεργαστές. Οι αρχιτεκτονικές που περιλαμβάνουν πολλαπλούς επεξεργαστές, διακρίνονται σε δυο κατηγορίες: α) τις αρχιτεκτονικές στενά συνδεδεμένων επεξεργαστών (tightly coupled architectures) και β) τις αρχιτεκτονικές χαλαρά συνδεδεμένων επεξεργαστών (loosely coupled architectures). Και στις δυο αρχιτεκτονικές παρουσιάζεται μια σειρά θεμάτων που θα πρέπει να αντιμετωπιστούν:

- Πώς να διαιρεθεί ένα πρόβλημα σε μικρότερα. Πιο συγκεκριμένα, πώς να αποσυντεθεί ένα πρόβλημα, έτσι ώστε μικρότερες εργασίες να εκτελεστούν παράλληλα.
- Πώς θα διασφαλιστεί η λήψη των δεδομένων που χρειάζεται κάθε μηχανή.
- Πώς να ανατεθούν εργασίες για εκτέλεση σε υπολογιστικούς κόμβους σε συστοιχίες μεγάλης κλίμακας, δεδομένου ότι ορισμένες μηχανές είναι πιο κατάλληλες από άλλες εξαιτίας των διαθέσιμων πόρων τους ή τοπικών περιορισμών που μπορεί να υπάρχουν.
- Πώς θα γίνει ο συντονισμός και ο συγχρονισμός μεταξύ των διαφόρων μηχανών.
- Πώς θα γίνει ο διαμοιρασμός των ενδιάμεσων αποτελεσμάτων από μια μηχανή σε όσες άλλες μηχανές τα χρειάζονται.
- Πώς αντιμετωπίζονται όλα τα παραπάνω θέματα στην παρουσία σφαλμάτων στο υλικό ή το λογισμικό.

1.3.1 Μοντέλα παραλληλίας σε επίπεδο εργασιών

Η πρώτη κατηγορία περιλαμβάνει μοντέλα που στοχεύουν στην επεξεργασία σε επίπεδο εργασιών (task parallel models) σε πολλαπλούς επεξεργαστές που έχουν μια κοινή μνήμη και τα δεδομένα μεταφέρονται μέσω ειδικών διαύλων επικοινωνίας. Οι στενά συνδεδεμένες αρχιτεκτονικές είναι συνήθως πιο αποδοτικές στην επεξεργασία, ενώ αποφεύγουν την αναπαραγωγή των δεδομένων και την άσκοπη μεταφορά τους στα διάφορα υποσυστήματά τους. Ωστόσο η όποια βελτίωση στις επιδόσεις και το μέγιστο μέγεθος των δεδομένων που μπορούν να αναλυθούν έχουν ένα άνω όριο. Σε τέτοιες περιπτώσεις, το κόστος επικοινωνίας καθίσταται απαγορευτικό με την αύξηση των επεξεργαστών. Παράλληλα η διαθέσιμη μνήμη κυμαίνεται στην τάξη των GB, οπότε εύκολα μπορεί να εξαντληθεί. Τέλος ένας παράγοντας που θα πρέπει να ληφθεί υπόψη είναι και το κόστος της αναβάθμισής του συστήματος, αφού για να επιτευχθεί κλιμάκωση στο σύστημα, χρειάζεται αντικατάσταση ολόκληρου του υλικού.

1.3.1.1 MPI και OpenMP

Στην πρώτη κατηγορία εντάσσονται τα παραδοσιακά μοντέλα παράλληλης επεξεργασίας όπως η Διεπαφή Μεταβίβασης Μηνυμάτων (Message Passing Interface – MPI) και η Ανοιχτή Πολυεπεξεργασία (Open Multi-Processing - OpenMP), που έχουν διαχρονικά αποδειχτεί πολύτιμοι σύμμαχοι στην παράλληλη επεξεργασία δεδομένων. Η Διεπαφή Μεταβίβασης Μηνυμάτων αποτελεί ένα σύνολο προδιαγραφών για την ανάπτυξη και τη χρήση βιβλιοθηκών μεταβίβασης μηνυμάτων που στοχεύει στη φορητότητα των εφαρμογών, στην υψηλή απόδοση και την επεκτασιμότητα. Υποστηρίζει επικοινωνίες σημείου προς σημείο αλλά και συλλογική επικοινωνία, ενώ παράλληλα ορίζει ρουτίνες για τη διαχείριση περιβάλλοντος, ομάδων διεργασιών και τοπολογιών. Προγράμματα που χρησιμοποιούν τη Διεπαφή Μεταβίβασης Μηνυμάτων εκτελούνται τόσο σε στενά όσο και σε χαλαρά διασυνδεδεμένες αρχιτεκτονικές. Από την άλλη, η Ανοιχτή Πολυεπεξεργασία είναι μια υλοποίηση του πολυνηματισμού (multithreading), μιας μεθόδου παραλληλίας, όπου ένα κύριο νήμα (δηλαδή μια σειρά εντολών που εκτελούνται σειριακά) διακλαδώνεται (forks) σε ένα πλήθος άλλων νημάτων και το σύστημα διαιρεί την προς εκτέλεση εργασία σε αυτά. Τα νήματα τρέχουν παράλληλα, με το υποκείμενο σύστημα εκτέλεσης να τα αναθέτει για εκτέλεση σε διαφορετικούς επεξεργαστές. Όταν τα νήματα ολοκληρώσουν την επεξεργασία, συγχρονίζονται (προαιρετικά) και ενώνονται με το κύριο νήμα. Και οι δυο προσεγγίσεις παρουσιάζουν πολλαπλά πλεονεκτήματα ορισμένα από τα οποία είναι η επίδοση (υπό συνθήκες) και η ικανότητα για κλιμάκωση (στην περίπτωση της Διεπαφής Μεταβίβασης Μηνυμάτων) αλλά και η φορητότητα του κώδικα για εκτέλεση σε διαφορετικά συστήματα. Ωστόσο η χρήση τους δυσχεραίνεται από πολλαπλά μειονεκτήματα όπως η δυσκολία στην ανάπτυξη εφαρμογών, η ευκολία στην παρουσία σφαλμάτων στα προγράμματα και ο σημαντικός επιπρόσθετος φόρτος εργασίας που χρειάζεται για την αντιμετώπιση αποτυχιών του υλικού.

1.3.2 Μοντέλα παραλληλίας σε επίπεδο δεδομένων

Η δεύτερη κατηγορία περιλαμβάνει μοντέλα που στοχεύουν στην επεξεργασία σε επίπεδο δεδομένων (data parallel models) και εκτελούνται σε αρχιτεκτονικές χαλαρά συνδεδεμένων επεξεργαστών. Σε τέτοιες αρχιτεκτονικές η υποδομή εκτέλεσης αποτελείται από συλλογές πολλαπλών επεξεργαστών, καθένας από τους οποίους έχει τη δική του τοπική μνήμη, συνήθως σε διαφορετικές τοποθεσίες όπου τα δεδομένα πρέπει να μεταφερθούν. Στις αρχιτεκτονικές αυτές τα υποσυστήματα των εφαρμογών είναι δυνατό να φιλοξενηθούν σε διαφορετικό υλικό, κάτι που καθιστά το όλο σύστημα ανθεκτικό στις αποτυχίες υλικού αλλά ταυτόχρονα και αποδοτικό, αφού οι επεξεργαστικές μονάδες μπορούν να επαναχρησιμοποιηθούν. Επίσης οι αναβαθμίσεις είναι εύκολες και σχετικά φθηνές, αφού ένας επεξεργαστικός κόμβος μπορεί να αντικατασταθεί ατομικά χωρίς να επηρεάζεται το υπόλοιπο σύστημα. Στον αντίποδα, μια εφαρμογή που εκτελείται σε μια χαλαρά διασυνδεδεμένη αρχιτεκτονική, χρειάζεται τη μετάδοση δεδομένων και την επικοινωνία μεταξύ των

κόμβων κάτι που εισάγει επιπρόσθετο φόρτο στην εφαρμογή, ειδικά όταν οι υπολογιστικοί κόμβοι είναι σε διαφορετικές γεωγραφικές περιοχές. Οι περιορισμοί των στενά συνδεδεμένων αρχιτεκτονικών έχουν στρέψει την προσοχή της ερευνητικής κοινότητας στην ανάπτυξη περιβαλλόντων εκτέλεσης και μοντέλων προγραμματισμού που προορίζονται για τις χαλαρά συνδεδεμένες αρχιτεκτονικές.

1.3.2.1 Μηχανές εκτέλεσης

Το Dryad [9] είναι ένα περιβάλλον εκτέλεσης παράλληλων εφαρμογών σε επίπεδο δεδομένων (data parallel runtime). Μια εφαρμογή στο Dryad, μοντελοποιείται σαν ένας κατευθυνόμενος ακυκλικός γράφος (directed acyclic graph), ο οποίος ορίζει τη ροή των δεδομένων της εφαρμογής και τις λειτουργίες που θα εκτελεστούν στα δεδομένα. Το Dryad εκτελεί τις διάφορες λειτουργίες στους κόμβους του γράφου τους οποίους το περιβάλλον εκτέλεσης κατανέμει σε πυρήνες (αν εκτελείται σε πολυπύρηνο σύστημα) ή σε υπολογιστές (αν εκτελείται σε συστοιχίες υπολογιστών). Ο χρονοπρογραμματισμός των υπολογιστικών κόμβων στο διαθέσιμο υλικό γίνεται από το περιβάλλον εκτέλεσης του Dryad, χωρίς καμία παρέμβαση από τον προγραμματιστή της εφαρμογής. Η ροή των δεδομένων μεταξύ των υπολογιστικών κόμβων γίνεται μέσω καναλιών επικοινωνίας, που είναι υλοποιημένα σε φυσικό επίπεδο με ροές TCP/IP, τη μνήμη ή προσωρινά αρχεία. Οι ροές χρησιμοποιούνται από το περιβάλλον εκτέλεσης για να μεταφέρουν ένα πεπερασμένο πλήθος δομημένων στοιχείων δεδομένων. Για τη δημιουργία ενός γράφου εκτέλεσης, το Dryad χρησιμοποιεί μια γλώσσα προγραμματισμού ειδικού σκοπού με το όνομα DryadLINQ [10].

Το μοντέλο της Απεικόνισης/Μείωσης (MapReduce) [11] περιλαμβάνει ένα απλό αλλά ταυτόχρονα αρκετά ισχυρό μοντέλο προγραμματισμού για την επεξεργασία δεδομένων μεγάλης κλίμακας. Παρόλο που υπάρχουν υλοποιήσεις του σε αρχιτεκτονικές μοιρασμένης μνήμης, στοχεύει πρωταρχικά σε συστοιχίες τυπικών υπολογιστικών συστημάτων για την επίτευξη παράλληλης και κατανεμημένης επεξεργασίας. Το μοντέλο Απεικόνισης/Μείωσης γρήγορα αναδείχθηκε ως το de facto πλαίσιο λογισμικού για την επεξεργασία και ανάλυση δεδομένων μεγάλης κλίμακας, εξαιτίας του απλού προγραμματιστικού μοντέλου και του αποδοτικού του συστήματος εκτέλεσης. Για το λόγο αυτό θα αναλυθεί διεξοδικότερα σε επόμενη παράγραφο.

Η διάδοση του μοντέλου Απεικόνισης/Μείωσης συνετέλεσε στη δημιουργία πολλαπλών επεκτάσεων και παραλλαγών τα τελευταία χρόνια. Τα νέα συστήματα αυτού του είδους μοιράζονται ένα αριθμό από κοινά χαρακτηριστικά, όπως το γεγονός ότι χρησιμοποιούν κάποιου είδους κατανεμημένο σύστημα, διαχειρίζονται ένα μεγάλο πλήθος διαδικασιών που μπορούν να θεωρηθούν απλά ως συναρτήσεις ορισμένες από τον προγραμματιστή οι οποίες τρέχουν παράλληλα και ότι υιοθετούν κάποιο μηχανισμό ο οποίος διαχειρίζεται τις αποτυχίες που συμβαίνουν κατά τη διάρκεια εκτέλεσης μιας εργασίας έτσι ώστε να μην είναι απαραίτητη η επανεκκίνηση του έργου από την αρχή. Οικογένειες τέτοιων συστημάτων είναι τα συστήματα ροής έργων (workflow systems) όπως τα

Clustera [12] και Hyracks [13] και οι επεκτάσεις αναδρομικής εκτέλεσης της Απεικόνισης/Μείωσης (recursive extensions to MapReduce).

Μια εναλλακτική προσέγγιση για το χειρισμό αποτυχιών κατά την υλοποίηση αναδρομικών αλγόριθμων, ακολουθείται στο υπολογιστικό μοντέλο Pregel [14]. Το Pregel επικεντρώνεται στην επεξεργασία γράφων μεγάλης κλίμακας, όπως δικτυακούς γράφους και γράφους κοινωνικής δικτύωσης. Μια υπολογιστική εργασία στο Pregel εκφράζεται με έναν κατευθυνόμενο γράφο που αποτελείται από κόμβους με κατευθυνόμενες ακμές. Τα προγράμματα εκφράζονται με ακολουθίες επαναλήψεων σε κάθε μια από τις οποίες ένας κόμβος λαμβάνει μηνύματα από την προηγούμενη επανάληψη, αλλάζει την κατάστασή του, στέλνει μηνύματα σε άλλους κόμβους και τροποποιεί την τοπολογία του γράφου. Αυτή η προσέγγιση γύρω από τους κόμβους του γράφου παρουσιάζει ευελιξία στην υλοποίηση ενός μεγάλου πλήθους αλγορίθμων. Το μοντέλο έχει σχεδιαστεί για αποδοτικές και κλιμακωτές υλοποιήσεις σε συστοιχίες τυπικών υπολογιστών, ενώ οι λεπτομέρειες της κατανομής των δεδομένων κρύβονται από τον προγραμματιστή.

Το Spark [15] [16] είναι ένα σχετικά νέο υπολογιστικό σύστημα που έχει λάβει αρκετή προσοχή τα τελευταία χρόνια. Το Spark είναι μια ταχεία μηχανή εκτέλεσης που βασίζεται σε μεγάλο βαθμό στη μνήμη. Το μοντέλο του Spark επιτρέπει στους υπολογιστικούς κόμβους να εκτελέσουν αποδοτικά αλγόριθμους που χρειάζονται γρήγορη επαναληπτική πρόσβαση σε σύνολα δεδομένων, όπως αυτούς της μηχανικής μάθησης. Η βελτίωση στις ταχύτητες οφείλεται στη διατήρηση των δεδομένων στη μνήμη καθ' όλη τη διάρκεια ενός έργου, αντί της επαναφόρτωσης των δεδομένων από το δίσκο κάθε φορά ξεχωριστά. Στον πυρήνα του Spark βρίσκεται η έννοια του ανθεκτικού κατανεμημένου συνόλου δεδομένων (resilient distributed dataset) που είναι μια αμετάβλητη συλλογή αντικειμένων τα οποία διαχωρίζονται και κατανέμονται για επεξεργασία σε πολλαπλούς υπολογιστικούς κόμβους σε μια συστοιχία υπολογιστών.

1.4 Το μοντέλο Απεικόνισης/Μείωσης

Το μοντέλο Απεικόνισης/Μείωσης (MapReduce) [11] είναι ένα μοντέλο παράλληλης επεξεργασίας που παρέχει τα μέσα για κλιμακούμενους υπολογισμούς με ανοχή σε σφάλματα. Αποτελείται από ένα μοντέλο προγραμματισμού και ένα υποκείμενο περιβάλλον εκτέλεσης και στοχεύει στην επεξεργασία δεδομένων μεγάλης κλίμακας χρησιμοποιώντας πολλαπλούς υπολογιστικούς κόμβους, σε συστοιχίες (αν όλοι οι κόμβοι βρίσκονται στο ίδιο τοπικό δίκτυο και αποτελούνται από παρόμοιο υλικό) ή πλέγματα (αν οι κόμβοι ανήκουν σε ανεξάρτητα κατανεμημένα συστήματα που βρίσκονται σε διαφορετικές γεωγραφικές περιοχές). Η επεξεργασία μπορεί να γίνει σε δεδομένα που είναι αποθηκευμένα σε ένα κατανεμημένο σύστημα αρχείων ή σε κάποια βάση δεδομένων. Κύριο χαρακτηριστικό του μοντέλου είναι ότι μπορεί να εκμεταλλευτεί την εντοπιότητα

των δεδομένων, δηλαδή να τα επεξεργαστεί στους κόμβους όπου είναι αποθηκευμένα, χωρίς να χρειάζεται η μετακίνησή τους μέσω του δικτύου.

Η δύναμη του μοντέλου Απεικόνισης/Μείωσης προκύπτει επίσης από την αφαιρετικότητα που επιτρέπει στους προγραμματιστές να εκμεταλλευτούν τη δύναμη μεγάλων συστοιχιών υπολογιστών χωρίς να χρειάζεται να ασχοληθούν με την πολυπλοκότητά του όλου εγχειρήματος. Με το μοντέλο της Απεικόνισης/Μείωσης ο προγραμματιστής επικεντρώνεται μόνο στη λογική της επεξεργασίας επί των δεδομένων, ενώ τις πολύπλοκες διεργασίες χαμηλού επιπέδου τις αναλαμβάνει εξ' ολοκλήρου το ίδιο το περιβάλλον εκτέλεσης με ένα κλιμακωτό, εύρωστο και αποδοτικό τρόπο.

Το μοντέλο Απεικόνισης/Μείωσης είναι χρήσιμο σε μια μεγάλη γκάμα προβλημάτων, μερικά από τα οποία είναι η κατανεμημένη αναζήτηση βάσει προτύπων, η ταξινόμηση, η εξαγωγή στατιστικών από αρχεία καταγραφής, η κατασκευή αντεστραμμένων ευρετηρίων, η συσταδοποίηση εγγράφων και η μηχανική μάθηση [17]. Επίσης πέραν των συστοιχιών υπολογιστών έχει μεταφερθεί και σε διάφορα άλλα υπολογιστικά περιβάλλοντα, όπως πολυπύρηνα συστήματα, πλέγματα υπολογιστών, συστήματα εθελοντικών υπολογιστών περιβάλλοντα υπολογιστικού νέφους και κινητά περιβάλλοντα.

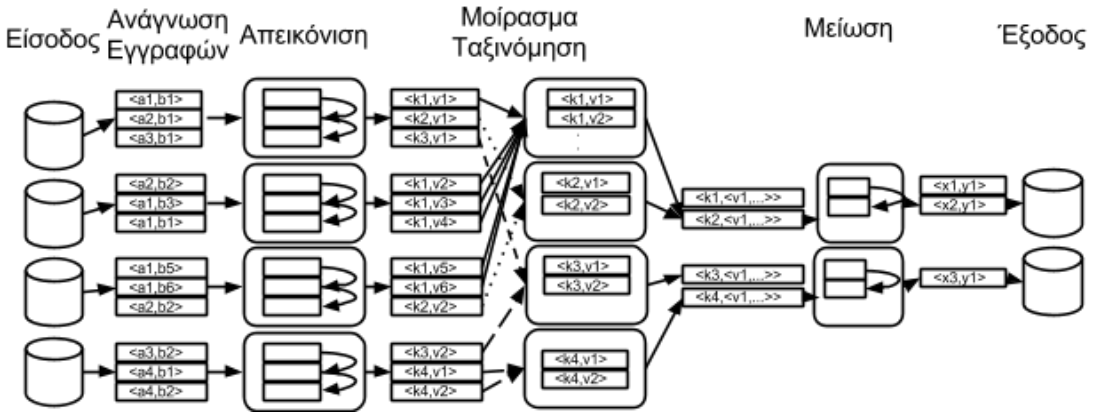
1.4.1 Το μοντέλο προγραμματισμού

Στο μοντέλο προγραμματισμού της Απεικόνισης/Μείωσης, οι μονάδες που ενσωματώνουν όλες τις επεξεργασίες επί των δεδομένων καλούνται έργα (jobs). Τα έργα αποτελούνται από τα δεδομένα εισόδου, τα συστατικά της Απεικόνισης/Μείωσης και τις διάφορες παραμέτρους εκτέλεσης, όπως για παράδειγμα πού βρίσκεται η είσοδος και που πρέπει να αποθηκεύεται η έξοδος. Σε ένα έργο Απεικόνισης/Μείωσης η επεξεργασία διαιρείται σε πολλαπλές εργασίες (tasks). Υπάρχουν δυο είδη εργασιών: α) οι εργασίες Απεικόνισης (Map) και β) οι εργασίες Μείωσης (Reduce). Τα δεδομένα αναπαρίστανται σαν ζευγάρια κλειδιού-τιμής (key-value pairs) και κάθε επεξεργασία που γίνεται στα πλαίσια ενός έργου, γίνεται πάνω σε αυτά.

Κάθε εργασία επεξεργάζεται ένα μικρό υποσύνολο των δεδομένων που της έχει ανατεθεί και έτσι ο συνολικός φόρτος εργασίας κατανέμεται σε όλη τη συστοιχία των υπολογιστικών κόμβων. Οι εργασίες Απεικόνισης γενικά εφαρμόζουν κάποιο είδος μετασχηματισμού ή φιλτραρίσματος, ενώ οι εργασίες Μείωσης δρουν επί των παραγόμενων δεδομένων και εφαρμόζουν κάποιο είδος άθροισης για την παραγωγή των τελικών αποτελεσμάτων. Τα ενδιάμεσα δεδομένα μεταξύ εργασιών Απεικόνισης και Μείωσης μεταφέρονται από τους υπολογιστικούς κόμβους που ολοκλήρωσαν τις εργασίες Απεικόνισης σε αυτούς που εκτελούν τις εργασίες Μείωσης για να τα ομαδοποιήσουν και να τα συναθροίσουν.

Το πλήρες μοντέλο προγραμματισμού της Απεικόνισης/Μείωσης φαίνεται στην Εικόνα 1.1. Η είσοδος σε ένα έργο Απεικόνισης/Μείωσης είναι ένα σύνολο αρχείων από το υποκείμενο κατανεμημένο σύστημα αρχείων. Τα αρχεία αυτά πρώτα χωρίζονται σε τμήματα (input splits)

σύμφωνα με κάποιους κανόνες ορισμένους από τον προγραμματιστή. Ένα τμήμα εισόδου αποτελεί το σύνολο δεδομένων που θα επεξεργαστεί μια εργασία Απεικόνισης.



Εικόνα 1.1 - Το μοντέλο προγραμματισμού Απεικόνισης/Μείωσης

Κάθε εργασία Απεικόνισης αποτελείται από τις εξής φάσεις: α) την ανάγνωση των εγγραφών ενός τμήματος εισόδου, β) τη διαδοχική εκτέλεση της συνάρτησης απεικόνισης σε κάθε εγγραφή ενός τμήματος εισόδου, γ) στο συνδυασμό των ζευγαριών εξόδου της απεικόνισης και στην κατάτμηση των ζευγαριών. Η έξοδος των εργασιών Απεικόνισης αποτελείται από ενδιάμεσα ζευγάρια κλειδιού-τιμής που στέλνονται στις εργασίες Μείωσης. Οι εργασίες Μείωσης αποτελούνται από τις εξής φάσεις: α) το μοίρασμα, β) την ταξινόμηση, γ) τη διαδοχική εκτέλεση της συνάρτησης μείωσης και δ) τη μορφοποίηση εξόδου. Οι κόμβοι όπου τρέχουν οι εργασίες Απεικόνισης στη βέλτιστη περίπτωση είναι αυτοί στους οποίους υπάρχουν τα δεδομένα, οπότε δεν χρειάζεται να μεταφερθούν δεδομένα στο δίκτυο. Τα συνιστάμενα μέρη ενός έργου Απεικόνισης/Μείωσης παρουσιάζονται στις επόμενες παραγράφους.

1.4.1.1 Ανάγνωση εγγραφών

Ο αναγνώστης εγγραφών (input reader) μεταφράζει ένα τμήμα εισόδου σε εγγραφές (ζευγάρια κλειδιού-τιμής). Ο σκοπός του αναγνώστη είναι η μετατροπή και όχι η επεξεργασία των εγγραφών. Ο αναγνώστης μεταφέρει τα δεδομένα στην Απεικόνιση σε μια μορφή κλειδιού/τιμής. Συνήθως το κλειδί του ζευγαριού αφορά τη θέση των δεδομένων στο αρχείο και η τιμή είναι το κομμάτι που απαρτίζει την εγγραφή. Η ανάγνωση των εγγραφών είναι μια εργασία στην οποία μπορεί να επέμβει ο προγραμματιστής.

1.4.1.2 Απεικόνιση

Η εργασία Απεικόνισης καλεί τη συνάρτηση απεικόνισης για κάθε ζευγάρι κλειδιού-τιμής του τμήματος δεδομένων που λαμβάνει. Η συνάρτηση απεικόνισης έχει την εξής μορφή:

```
απεικόνιση(k1,v1)  →  λίστα(k2,v2)
```

όπου k1 το κλειδί της εγγραφής εισόδου, v1 η τιμή της εγγραφής εισόδου και λίστα(k2,v2) μια λίστα από ζευγάρια κλειδιού-τιμής εξόδου. Η επιλογή του τύπου του κλειδιού και της τιμής τόσο της εισόδου όσο και της εξόδου είναι πολύ σημαντική για το σκοπό της επεξεργασίας. Το κλειδί και η τιμή εξόδου της συνάρτησης απεικόνισης είναι συνήθως διαφορετικά από αυτά της εισόδου, αλλά κάτι τέτοιο δεν είναι απαραίτητο. Τα ζευγάρια κλειδιού-τιμής εξόδου αποθηκεύονται προσωρινά στη μνήμη του υπολογιστικού κόμβου και περιοδικά αποθηκεύονται στον τοπικό σκληρό δίσκο. Η αποθήκευση αυτή γίνεται με τέτοιο τρόπο ώστε τα ζευγάρια με το ίδιο κλειδί να ομαδοποιούνται μαζί. Η διαδικασία αυτή γίνεται συνήθως με μια συνάρτηση κατακερματισμού, ωστόσο είναι δυνατόν να χρησιμοποιηθεί οποιαδήποτε ντετερμινιστική συνάρτηση. Γι' αυτήν την εργασία λοιπόν ο προγραμματιστής θα πρέπει να ορίσει την επεξεργασία που θα γίνεται μέσα στη συνάρτηση της απεικόνισης.

1.4.1.3 Συνδυασμός

Ο συνδυαστής (combiner), είναι μια προαιρετική βελτιστοποίηση που ομαδοποιεί τα δεδομένα στη φάση της Απεικόνισης. Λαμβάνει τα ενδιάμεσα κλειδιά που προέκυψαν από την Απεικόνιση και εφαρμόζει μια συνάρτηση ορισμένη από τον προγραμματιστή που αθροίζει τιμές μέσα στο εύρος των δεδομένων που επεξεργάζεται η εργασία της Απεικόνισης. Σε πολλές περιπτώσεις κάτι τέτοιο μειώνει το πλήθος των δεδομένων που πρέπει να μεταφερθούν μέσα από το δίκτυο.

1.4.1.4 Διαχωρισμός

Ο διαχωριστής (partitioner) λαμβάνει τα ενδιάμεσα ζευγάρια κλειδιού-τιμής από την Απεικόνιση (ή το Συνδυαστή αν χρησιμοποιείται) και τα διαχωρίζει σε κομμάτια (shards), ένα για κάθε εργασία Μείωσης. Η πιο συνηθισμένη προσέγγιση είναι να υπολογίζεται ο κώδικας κατακερματισμού του κλειδιού και στη συνέχεια να υπολογίζεται το υπόλοιπο της διαίρεσης με το πλήθος των εργασιών Μείωσης. Έτσι διασφαλίζεται η τυχαία κατανομή των κλειδιών με ομοιόμορφο τρόπο, αλλά ταυτόχρονα διασφαλίζει ότι τα ίδια κλειδιά με διαφορετικές τιμές σε διαφορετικούς υπολογιστικούς κόμβους, θα καταλήξουν στην ίδια εργασία Μείωσης. Η διαδικασία του διαχωρισμού μπορεί να οριστεί από τον προγραμματιστή. Τα δεδομένα που έχουν ήδη διαχωριστεί γράφονται στο τοπικό σύστημα αρχείων για κάθε εργασία Απεικόνισης και περιμένουν μέχρι να μεταφερθούν στην αντίστοιχη εργασία Μείωσης.

1.4.1.5 Μοίρασμα και ταξινόμηση

Η εργασία της Μείωσης ξεκινά με το βήμα του μοιράσματος και της ταξινόμησης. Αυτό το βήμα λαμβάνει σαν είσοδο τα αρχεία εξόδου όλων των διαχωριστών και τα μεταφέρει στους αντίστοιχους υπολογιστικούς κόμβους που τρέχουν τις εργασίες των Μειώσεων. Αυτά τα κομμάτια δεδομένων ταξινομούνται κατά κλειδί σε μια μεγαλύτερη λίστα δεδομένων. Ο σκοπός αυτής της ταξινόμησης είναι να ομαδοποιηθούν τα ίδια κλειδιά μαζί έτσι ώστε οι τιμές τους να προσπελαστούν ευκολότερα στη συνάρτηση μείωσης που θα εκτελεστεί στη συνέχεια. Αυτή η φάση δεν γίνεται να αλλαχθεί από

τον προγραμματιστή αφού την αναλαμβάνει το περιβάλλον εκτέλεσης. Η μόνη παρέμβαση που μπορεί να κάνει ο προγραμματιστής είναι στον τρόπο με τον οποίο θα γίνεται η σύγκριση των κλειδιών κατά την ταξινόμηση.

1.4.1.6 Μείωση

Η εργασία Μείωσης (Reduce) λαμβάνει τα ομαδοποιημένα δεδομένα σαν είσοδο και εκτελεί τη συνάρτηση μείωσης για κάθε κλειδί-ομάδα. Η συνάρτηση μείωσης έχει την εξής μορφή:

```
μείωση(k2,λίστα(v2))        →        λίστα(k3,v3)
```

όπου k2, το κλειδί εξόδου της συνάρτησης μείωσης και v2 οι τιμές που αντιστοιχούν σε αυτό το κλειδί από όλες τις εργασίες Απεικόνισης. Το τελικό αποτέλεσμα είναι μια λίστα από ζευγάρια κλειδιού-τιμής οι τύποι των οποίων μπορεί να είναι διαφορετικοί από τους αντίστοιχους της εισόδου. Η συνάρτηση δέχεται σαν είσοδο το κλειδί της ομάδας και τις τιμές της που σχετίζονται με αυτό το κλειδί. Σε αυτή τη συνάρτηση μπορούν να λάβουν χώρα διαφόρων ειδών επεξεργασίες, όπως συναθροίσεις ή φιλτραρίσματα. Κατά την ολοκλήρωση της, η συνάρτηση μπορεί να έχει σαν έξοδο κανένα ή περισσότερα ζευγάρια τα οποία, στέλνει στο τελικό βήμα που είναι η μορφοποίηση της εξόδου. Όπως και η συνάρτηση απεικόνισης, η συνάρτηση μείωσης ορίζεται από τον προγραμματιστή.

1.4.1.7 Μορφοποίηση εξόδου

Η μορφοποίηση εξόδου (output format) μετατρέπει το κλειδί-τιμή από την έξοδο της συνάρτησης μείωσης στην τελική μορφή που θα έχει κατά την αποθήκευσή της στο σύστημα αρχείων. Μια τέτοια μορφή συνήθως διαχωρίζει το κλειδί από την τιμή αλλά και από τα υπόλοιπα ζευγάρια κλειδιού-τιμής. Η μορφή αυτή προφανώς μπορεί να οριστεί κατάλληλα από τον προγραμματιστή ανάλογα με τις απαιτήσεις του προγράμματος.

1.4.2 Το υποκείμενο περιβάλλον εκτέλεσης

Ο διαχωρισμός του μοντέλου προγραμματισμού από το περιβάλλον εκτέλεσης στο μοντέλο Απεικόνισης/Μείωσης καθιστά δυνατή την υλοποίηση προγραμμάτων και την εκτέλεσή τους σε διαφορετικά περιβάλλοντα. Το μοντέλο Απεικόνισης/Μείωσης ωστόσο, έχει επικρατήσει να συνδέεται με χρήση κατανεμημένων περιβαλλόντων εκτέλεσης και πιο συγκεκριμένα συστοιχίες κοινών υπολογιστών. Το υποκείμενο περιβάλλον εκτέλεσης αναλαμβάνει τη διαχείριση της συστοιχίας και την επικοινωνία μεταξύ των υπολογιστικών κόμβων, για να φέρει σε πέρας την εκτέλεση των έργων Απεικόνισης/Μείωσης. Ένα έργο Απεικόνισης/Μείωσης λοιπόν, έτσι όπως έχει αναπτυχθεί από τον προγραμματιστή, υποβάλλεται σε έναν κεντρικό κόμβο της συστοιχίας υπολογιστών και το περιβάλλον εκτέλεσης αναλαμβάνει όλα τα υπόλοιπα. Ορισμένες από τις αρμοδιότητες του περιβάλλοντος εκτέλεσης ακολουθούν στις επόμενες παραγράφους.

1.4.2.1 Χρονοπρογραμματισμός

Σε μεγάλα έργα Απεικόνισης/Μείωσης το συνολικό πλήθος εργασιών που πρέπει να εκτελεστούν μπορεί να είναι πολύ μεγαλύτερο από το πλήθος των υπολογιστικών κόμβων της συστοιχίας όπου εκτελείται το έργο. Έτσι θα πρέπει να διατηρηθεί μια ουρά προτεραιοτήτων που θα αναθέτει υπολογιστικούς κόμβους, όταν αυτοί γίνουν διαθέσιμοι, στις εργασίες με τη μεγαλύτερη αναμονή. Άλλο ένα θέμα που προκύπτει στον χρονοπρογραμματισμό είναι ο συντονισμός των εργασιών που ανήκουν σε διαφορετικά έργα που εκτελούνται ταυτόχρονα στην ίδια συστοιχία [18]. Ένα σημαντικό θέμα που προκύπτει επίσης αφορά τις εργασίες που απαιτούν περισσότερο χρόνο από τις υπόλοιπες για να ολοκληρωθούν, για παράδειγμα λόγω κάποιου σφάλματος στο υλικό. Μια αντιμετώπιση του προβλήματος αυτού είναι η υποθετική εκτέλεση (speculative execution), με την οποία το περιβάλλον εκτέλεσης μεταφέρει ένα αντίγραφο της εργασίας αυτής σε κάποιον από τους κόμβους με τα εφεδρικά αντίγραφα των τμημάτων των δεδομένων για να ολοκληρωθεί εκεί. Με την προσέγγιση αυτή ωστόσο, δεν μπορεί να αντιμετωπιστεί μια άλλη αιτία εμφάνισης καθυστερούμενων εργασιών που είναι η ασυμμετρία (skew) στην κατανομή των ενδιάμεσων τιμών.

1.4.2.2 Εντοπιότητα δεδομένων

Μια από τις βασικότερες ιδέες της Απεικόνισης/Μείωσης είναι η μεταφορά του κώδικα στα δεδομένα για την εκτέλεση των απαιτούμενων εργασιών. Η μεταφορά του προγράμματος στους κόμβους όπου υπάρχουν τα δεδομένα προς ανάλυση εξαρτάται από την υλοποίηση του κατανεμημένου συστήματος αρχείων. Ο χρονοπρογραμματιστής επιχειρεί να ξεκινήσει μια εργασία στον κόμβο όπου είναι αποθηκευμένο το συγκεκριμένο μπλοκ δεδομένων που αυτή χρειάζεται. Αν ο κόμβος είναι ήδη προγραμματισμένο να εκτελέσει άλλες εργασίες, οι νέες εργασίες θα ξεκινήσουν σε εναλλακτικούς κόμβους και αυτό μπορεί να έχει σαν συνέπεια τη μεταφορά δεδομένων. Μια σημαντική βελτιστοποίηση που εφαρμόζεται σε αυτή τη διαδικασία είναι να προτιμώνται κόμβοι που βρίσκονται στο ίδιο ράφι (rack) με τον κόμβο που περιέχει το υπό επεξεργασία κομμάτι δεδομένων, αφού το εύρος ζώνης εντός του ραφιού είναι μεγαλύτερο από αυτό εκτός του.

1.4.2.3 Συγχρονισμός

Στο μοντέλο Απεικόνισης/Μείωσης το μοντέλο επικοινωνίας είναι μονόδρομο και συμβαίνει μόνο μεταξύ των εργασιών Απεικόνισης και των εργασιών Μείωσης. Οι δυο αυτές φάσεις είναι ξεχωριστές και η μετάδοση των δεδομένων από την πρώτη στη δεύτερη γίνεται με τη διαδικασία του μοιράσματος και της ταξινόμησης. Στη διαδικασία αυτή, τα ενδιάμεσα δεδομένα αντιγράφονται από τους κόμβους των Απεικονίσεων στους κόμβους των Μειώσεων και μόνο μετά την ολοκλήρωσή της μπορεί να ξεκινήσει την εκτέλεση μια διαδικασία Μείωσης.

1.4.2.4 Διαχείριση σφαλμάτων

Το μοντέλο Απεικόνισης/Μείωσης εφόσον στοχεύει να εκτελείται σε συστοιχίες τυπικών υπολογιστών του εμπορίου, συνήθως χαμηλού κόστους, θεωρεί δεδομένο ότι θα συμβαίνουν συχνά σφάλματα στο υλικό και στο λογισμικό κατά την εκτέλεση των έργων. Επίσης τα υπολογιστικά κέντρα υπόκεινται σε προγραμματισμένες (π.χ. για συντήρηση υλικού ή αναβάθμιση λογισμικού) αλλά και ξαφνικές διακοπές στις υπηρεσίες τους (π.χ. διακοπή παροχής ρεύματος ή διακοπή συνδεσιμότητας). Σφάλματα επίσης μπορούν να συμβούν εξαιτίας του λογισμικού. Η διαχείριση των εξαιρέσεων σε ένα πρόγραμμα πρέπει να γίνεται αποδοτικά ώστε να ανακύπτει χωρίς προβλήματα και να συνεχίζει την εκτέλεσή του. Στα προβλήματα μεγάλου εύρους και η παραμικρή λεπτομέρεια μπορεί να αποβεί ολέθρια στην εκτέλεση ενός έργου. Επιπλέον, κάθε επαρκώς μεγάλο σύνολο δεδομένων έχει αυξημένη πιθανότητα να περιέχει αλλοιωμένες εγγραφές πέραν από τις προβλέψεις των προγραμματιστών κάτι που οδηγεί σε απρόβλεπτα σφάλματα. Κάθε περιβάλλον εκτέλεσης Απεικόνισης/Μείωσης πρέπει να ανταποκρίνεται επαρκώς σε τέτοια ζητήματα.

1.4.2.5 Το κατανεμημένο σύστημα αρχείων

Εφόσον το μοντέλο Απεικόνισης/Μείωσης στοχεύει στην εκτέλεση σε συστοιχίες υπολογιστών, τα δεδομένα που επεξεργάζεται θα πρέπει να κατανέμονται σε πολλαπλές μηχανές. Τα συστήματα αρχείων που διαχειρίζονται την αποθήκευση αρχείων σε δίκτυα υπολογιστών καλούνται κατανεμημένα συστήματα αρχείων υπολογιστών (distributed file systems). Η έννοια των κατανεμημένων συστημάτων αρχείων δεν είναι καινούργια [19] [20] [21] ωστόσο το κατανεμημένο σύστημα αρχείων του μοντέλου Απεικόνισης/Μείωσης διαφέρει από προηγούμενες αρχιτεκτονικές σε συγκεκριμένα σημεία αφού στοχεύει ειδικά σε μεγάλα φορτία επεξεργασίας δεδομένων [22].

Το κατανεμημένο σύστημα αρχείων της Απεικόνισης/Μείωσης είναι σχεδιασμένο να αποθηκεύει μεγάλα αρχεία σε τυπικούς υπολογιστικούς κόμβους και να παρέχει πρόσβαση σε αυτά με συνεχόμενη προσπέλαση. Το μέγεθος των αρχείων μπορεί να κυμαίνεται από εκατοντάδες MB μέχρι εκατοντάδες TB, ωστόσο δεν είναι αποδοτικό να αποθηκεύονται πολλά μικρά αρχεία. Τα δεδομένα διαιρούνται σε μπλοκ και αντιγράφονται στους τοπικούς κόμβους της συστοιχίας. Επίσης το σύστημα δεν υποστηρίζει την ενημέρωση των αρχείων ούτε τη χαμηλή καθυστέρηση στη μεταφορά τους καθώς έχει προτιμηθεί η βελτιστοποίηση της ανάγνωσης των αρχείων μετά την εισαγωγή τους στο σύστημα και ο ρυθμός εξυπηρέτησης. Τέλος, το κατανεμημένο σύστημα αρχείων δεν απαιτεί τη χρήση ακριβού υλικού, αφού το σύστημα είναι σχεδιασμένο να τρέχει πάνω σε συστοιχίες τυπικών υπολογιστών του εμπορίου.

Το κατανεμημένο σύστημα αρχείων συντονίζεται από ένα κεντρικό κόμβο. Ο κόμβος αυτός συντηρεί τον χώρο ονομάτων των αρχείων και πιο συγκεκριμένα διατηρεί τα μεταδεδομένα, τη δομή των δεδομένων, την αντιστοίχηση μπλοκ/αρχείου, την τοποθεσία των μπλοκ και τις δυνατότητες πρόσβασης στα αρχεία αυτά. Οι κατά τόπους κόμβοι διαχειρίζονται τα πραγματικά μπλοκ

δεδομένων. Για κάθε μπλοκ δεδομένων δημιουργούνται πλεονάζοντα αντίγραφα για να διασφαλιστεί η διαθεσιμότητα των δεδομένων και η αξιοπιστία του συστήματος. Τα αντίγραφα αυτά συνήθως μεταφέρονται και σε διαφορετικά ράφια για να αποφευχθεί η απώλεια των δεδομένων με μια πιθανή αποτυχία ενός κόμβου, του δικτυακού εξοπλισμού ενός ραφιού αλλά και για να αυξηθεί η πιθανότητα να μεταφερθεί κάποια εργασία απεικόνισης στον κόμβο κατά το χρονοπρογραμματισμό ενός έργου. Ο κόμβος ονομάτων επικοινωνεί περιοδικά με κάθε κόμβο για να διασφαλίσει τη διαθεσιμότητα των αντιγράφων στους κόμβους: αν δεν υπάρχουν αρκετά αντίγραφα, ο κόμβος ονομάτων δημιουργεί νέα. Αν υπάρχουν περισσότερα αντίγραφα από όσα πρέπει, τα περιττά αντίγραφα διαγράφονται.

Συνοψίζοντας, ο κεντρικός κόμβος του συστήματος αρχείων α) διαχειρίζεται τα αρχεία του κατανεμημένου συστήματος, β) συντονίζει τις λειτουργίες επί των αρχείων και γ) συντηρεί την υγεία του συστήματος αρχείων. Από τις τρεις αυτές αρμοδιότητες γίνεται ξεκάθαρη η σημασία του κόμβου αυτού για το όλο σύστημα αφού αν για κάποιο λόγο γίνει μη διαθέσιμος, ολόκληρο το σύστημα αρχείων αλλά και όλα τα έργα Απεικόνισης/Μείωσης που εκτελούνται στη συστοιχία θα καταρρεύσουν. Αυτή η αδυναμία είναι σχετικά μικρή αφού η φύση των λειτουργιών που αυτός εκτελεί δεν είναι ιδιαίτερα απαιτητική. Επειδή δεν μεταφέρονται δεδομένα από τα αρχεία μέσω του κεντρικού κόμβου, αφού μεταφέρονται κατευθείαν από τους κόμβους δεδομένων, σπάνια συμβαίνει συμφόρηση στον κεντρικό κόμβο. Ωστόσο για περισσότερη ασφάλεια, η συνηθισμένη πρακτική είναι να υπάρχει ένας εφεδρικός κεντρικός κόμβος σε περίπτωση απώλειας του πρώτου.

1.4.3 Πλεονεκτήματα και μειονεκτήματα

Από τις προηγούμενες παραγράφους κατέστη σαφές ότι η ισχύς του μοντέλου Απεικόνισης/Μείωσης οφείλεται στην έμφυτη κλιμάκωση που παρουσιάζουν όσα προγράμματα υλοποιούνται σε αυτό. Το μοντέλο παραλληλοποιεί αυτόματα τους υπολογισμούς στους υπολογιστικούς κόμβους ανεξαρτήτως του μεγέθους των δεδομένων εισόδου. Όλες οι λεπτομέρειες που αφορούν τον ταυτοχρονισμό, τη μεταφορά δεδομένων μεταξύ των κόμβων και η εκτέλεση των προγραμμάτων τακτοποιούνται από το περιβάλλον εκτέλεσης.

Το μοντέλο Απεικόνισης/Μείωσης είναι απλό και εύκολο στη χρήση. Ένας προγραμματιστής μπορεί να ορίσει ένα έργο μόνο με συναρτήσεις Απεικόνισης και Μείωσης, χωρίς να χρειάζεται να οριστεί πώς θα κατανεμηθούν οι εργασίες στους κόμβους. Επίσης το μοντέλο επειδή δεν απαιτεί τον ορισμό κάποιου αυστηρού σχήματος για τα δεδομένα, είναι πιο ευέλικτο στην επεξεργασία αδόμητων δεδομένων. Η Απεικόνιση/Μείωση επίσης είναι ανεξάρτητη του συστήματος αποθήκευσης των δεδομένων, επιτρέποντας έτσι το διάβασμα τόσο από ένα κατανεμημένο σύστημα αρχείων όσο και από μια κατανεμημένη βάση δεδομένων.

Από την άλλη το μοντέλο Απεικόνισης/Μείωσης στερείται ορισμένων χαρακτηριστικών που έχουν αποδειχθεί υψίστης σημασίας κατά την ανάλυση δεδομένων από σχεσιακά συστήματα βάσεων δεδομένων. Το μοντέλο δεν διαθέτει κάποια γλώσσα υψηλού επιπέδου για την υποβολή ερωτημάτων

πάνω στα δεδομένα και για το λόγο αυτό χρειάζεται η ανάπτυξη ολοκληρωμένων προγραμμάτων Απεικόνισης/Μείωσης. Το μοντέλο επίσης δεν χρησιμοποιεί κάποιο σχήμα για τα δεδομένα και για το λόγο αυτό κάθε πρόγραμμα θα πρέπει υποχρεωτικά να τα φορτώσει και να τα μετατρέψει σε μια πιο κατάλληλη μορφή για ανάλυση, κάτι που μπορεί να επιφέρει αχρείαστο φόρτο. Η Απεικόνιση/Μείωση ναι μεν παρέχει ένα προγραμματιστικό μοντέλο που είναι απλό, αλλά το μοντέλο αυτό έχει μια προκαθορισμένη ροή που μπορούν να ακολουθήσουν τα δεδομένα. Κατά συνέπεια πολλοί πολύπλοκοι αλγόριθμοι είναι δύσκολο να υλοποιηθούν σε ένα μόνο έργο Απεικόνισης/Μείωσης. Επιπρόσθετα το μοντέλο είναι σχεδιασμένο να δέχεται μόνο ένα σύνολο δεδομένων σαν είσοδο, οπότε αλγόριθμοι που χρειάζονται περισσότερες της μιας εισόδου δεν υποστηρίζονται. Τέλος το μοντέλο δεν έχει σχέδια εκτέλεσης για τη βελτιστοποίηση της εκτέλεσης των έργων κάτι που μπορεί να δυσχεράνει την επίδοση των έργων.

1.5 Το οικοσύστημα εφαρμογών για Μεγάλα Δεδομένα

Τα τελευταία χρόνια, η βιομηχανία και η ερευνητική κοινότητα έχουν συνεισφέρει στην γκάμα των διαθέσιμων εργαλείων για την επεξεργασία Μεγάλων Δεδομένων μια πληθώρα εφαρμογών. Οι εφαρμογές αυτές ανήκουν σε πέντε κατηγορίες.

Στην πρώτη κατηγορία ανήκουν τα υπολογιστικά συστήματα μαζικών επεξεργασιών (batch computation systems). Τέτοια συστήματα έχουν μεγάλο ρυθμό εκτέλεσης εργασιών (throughput) αλλά ταυτόχρονα έχουν μεγάλη καθυστέρηση (latency), κάτι που σημαίνει ότι παρόλο που μπορεί να επιτελούν απλές εργασίες, χρειάζεται πολύς χρόνος για να τις ολοκληρώσουν. Το πιο γνωστό σύστημα σε αυτή την κατηγορία είναι το Hadoop [23] που αποτελεί υλοποίηση του μοντέλου Απεικόνισης/Μείωσης.

Στη δεύτερη κατηγορία ανήκουν τα πλαίσια σειριοποίησης (serialization frameworks). Τα πλαίσια σειριοποίησης είναι σύνολα εργαλείων και βιβλιοθηκών προγραμματισμού που επιτρέπουν τη μεταφορά δεδομένων μεταξύ ετερογενών εφαρμογών, συχνά υλοποιημένων σε διαφορετικές γλώσσες προγραμματισμού. Μπορούν να μετατρέψουν τα δεδομένα σε μια σειρά από bytes από οποιαδήποτε γλώσσα και να τα επαναφέρουν στην αρχική τους μορφή πάλι από οποιαδήποτε γλώσσα. Τα πλαίσια σειριοποίησης ορίζουν μια Γλώσσα Ορισμού Σχήματος (Schema Definition Language) για τον ορισμό της μορφής των δεδομένων (συχνά αντιμετωπίζονται ως αντικείμενα) και παρέχουν μηχανισμούς χειρισμού διαφορετικών εκδόσεων έτσι ώστε τα σχήματα να μπορούν να εξελιχθούν χωρίς να δημιουργούνται προβλήματα με τα υπάρχοντα αντικείμενα. Τα πιο γνωστά πλαίσια σειριοποίησης είναι τα Thrift [24], Protocol Buffers [25] και Avro [26].

Στην τρίτη κατηγορία ανήκουν οι Βάσεις δεδομένων NoSQL. Οι βάσεις δεδομένων NoSQL παρέχουν μηχανισμούς για την αποθήκευση και την ανάκτηση δεδομένων που μοντελοποιούνται διαφορετικά από το παραδοσιακό σχεσιακό μοντέλο. Οι λόγοι που οδηγούν στη χρήση τέτοιων

βάσεων περιλαμβάνουν την απλούστευση του σχεδιασμού τους, την οριζόντια κλιμάκωση και τον πλήρη έλεγχο στη διαθεσιμότητα των δεδομένων. Οι δομές δεδομένων που χρησιμοποιούνται από τις βάσεις NoSQL (π.χ. κλειδιά-τιμές, γράφοι ή κείμενα) διαφέρουν από αυτές που χρησιμοποιούνται σε σχεσιακές βάσεις δεδομένων, κάτι που τις καθιστά σε ορισμένες περιπτώσεις γρηγορότερες από τις αντίστοιχες σχεσιακές. Παραδείγματα τέτοιων βάσεων είναι τα Cassandra [27], HBase [28], MongoDB [29] και CouchDB [30]. Η καταλληλότητα μιας συγκεκριμένης NoSQL βάσης δεδομένων εξαρτάται από το πρόβλημα προς επίλυση.

Στην τέταρτη κατηγορία ανήκουν τα συστήματα ουρών και μηνυμάτων (queuing/messaging systems). Ένα σύστημα ουρών και μηνυμάτων παρέχει έναν τρόπο αποστολής και λήψης μηνυμάτων μεταξύ διεργασιών με έναν ανεκτικό σε σφάλματα και ασύγχρονο τρόπο. Μια ουρά μηνυμάτων είναι ένα κύριο συστατικό της επεξεργασίας σε πραγματικό χρόνο. Παράδειγμα συστήματος σε αυτήν την κατηγορία είναι το Apache Kafka [31].

Τέλος, στην πέμπτη κατηγορία ανήκουν τα συστήματα επεξεργασίας πραγματικού χρόνου (realtime computation systems). Τα συστήματα επεξεργασίας πραγματικού χρόνου, είναι συστήματα επεξεργασίας ροών δεδομένων με μεγάλο ρυθμό εκτέλεσης εργασιών και χαμηλή καθυστέρηση. Δεν μπορούν να πραγματοποιήσουν το εύρος της επεξεργασίας των συστημάτων μαζικών επεξεργασιών, αλλά επεξεργάζονται τα δεδομένα σε μεγάλες ταχύτητες. Ένα παράδειγμα τέτοιου συστήματος είναι το Storm [32].

Κεφάλαιο 2
Επαγωγή κανόνων κατηγοριοποίησης με την τεχνική Απεικόνισης/Μείωσης

2.1 Εισαγωγή

Στη σύγχρονη εποχή, λόγω της συνεχούς παραγωγής δεδομένων από διάφορες, συνήθως ετερογενείς πηγές, η ανάγκη για κλιμακωτή επεξεργασία δεδομένων πολύ μεγάλου μεγέθους έχει γίνει επιτακτική. Το πρότυπο Απεικόνισης/Μείωσης γρήγορα αναδείχθηκε ως το de facto πλαίσιο λογισμικού για την επεξεργασία και ανάλυση δεδομένων μεγάλης κλίμακας, εξαιτίας του απλού προγραμματιστικού μοντέλου και του αποδοτικού του συστήματος εκτέλεσης. Ωστόσο αυτή η απλότητα συνοδεύεται από ένα σημαντικό κόστος: το μονόδρομο μοντέλο επικοινωνίας του και την έλλειψη εγγενούς και αποδοτικής υποστήριξης επαναληπτικών εκτελέσεων διαδικασιών ορισμένων από τον προγραμματιστή. Το γεγονός αυτό περιορίζει την εφαρμογή του συγκεκριμένου μοντέλου σε πεδία όπως αυτό της μηχανικής μάθησης.

Σε αυτό το κεφάλαιο περιγράφεται η υλοποίηση ενός αλγόριθμου επαγωγής κανόνων για κατηγοριοποίηση δεδομένων που βασίζεται στο μοντέλο Απεικόνισης/Μείωσης, με στόχο την τελική κατασκευή ενός αποδοτικού μοντέλου κατηγοριοποίησης σε όσο το δυνατόν λιγότερες επαναλήψεις. Επίσης περιγράφεται η υλοποίηση μιας προσέγγισης διακριτοποίησης ως ένα βήμα προεπεξεργασίας των δεδομένων το οποίο καθιστά δυνατή την επαγωγή κανόνων από αυτά. Μετά από μια αναλυτική περιγραφή των δυο προσεγγίσεων, η προτεινόμενη λύση αξιολογείται από τρεις οπτικές γωνίες: α) την ακρίβειά της ως προς την κατηγοριοποίηση, β) την επίδοσή της ως προς την παράλληλη εκτέλεση και γ) το κόστος στην επικοινωνία μεταξύ των επεξεργαστικών μονάδων. Η αξιολόγηση υποδεικνύει ότι η προσέγγιση είναι κλιμακωτή και επειδή παράγει μοντέλα εύκολα κατανοητά στους ανθρώπους, μπορεί να αποδειχθεί χρήσιμη σε μια πλειάδα εφαρμογών.

2.2 Το πρόβλημα της επαγωγής κανόνων κατηγοριοποίησης

Το πρόβλημα της επαγωγής κανόνων κατηγοριοποίησης μελετά την κατασκευή ενός συνόλου κανόνων από ένα σύνολο ήδη κατηγοριοποιημένων παραδειγμάτων, που μπορούν να χρησιμοποιηθούν για την κατηγοριοποίηση νέων, άγνωστων στιγμιότυπων. Ένας πιο αυστηρός ορισμός του προβλήματος επαγωγής κανόνων κατηγοριοποίησης παρέχεται στο βιβλίο [33]:

Με δεδομένα τα εξής:

- Μιας γλώσσας περιγραφής δεδομένων, που περιγράφει τη μορφή τους,
- Μιας γλώσσας περιγραφής υποθέσεων, που περιγράφει τη μορφή των κανόνων
- Μιας συνάρτησης κάλυψης, που ορίζει αν ένας κανόνας καλύπτει ένα παράδειγμα
- Μιας ιδιότητας που αναπαριστά την κατηγορία
- Ενός συνόλου παραδειγμάτων στη μορφή που περιγράφεται από τη γλώσσα περιγραφής δεδομένων

Στόχος είναι η εύρεση μιας υπόθεσης στη μορφή ενός συνόλου κανόνων, όπως περιγράφεται από τη γλώσσα περιγραφής υποθέσεων που έχει υιοθετηθεί, η οποία θα είναι:

- Πλήρης, δηλαδή καλύπτει όλα τα παραδείγματα, και
- Συνεπής, δηλαδή προβλέπει τη σωστή κατηγορία για όλα τα παραδείγματα

Το σύνολο κανόνων θα πρέπει να αποκαλύπτει την κρυμμένη σχέση μεταξύ των ιδιοτήτων εισόδου και της κατηγορίας και θα πρέπει να γενικεύει αυτή τη σχέση και σε νέα, άγνωστα στιγμιότυπα. Στη συνέχεια περιγράφονται καθένα από τα συστατικά του προβλήματος.

2.2.1 Αναπαράσταση δεδομένων

Η είσοδος σε έναν αλγόριθμο επαγωγής κανόνων αποτελείται από ένα σύνολο παραδειγμάτων εκπαίδευσης, δηλαδή στιγμιότυπων των οποίων η κατηγορία είναι γνωστή και σαφώς προσδιορισμένη. Συνήθως τα στιγμιότυπα αυτά περιγράφονται με μια αναπαράσταση ιδιότητας-τιμής: Η περιγραφή ενός στιγμιότυπου έχει τη μορφή μιας πλειάδας $(v_{1j}, v_{2j}, \dots, v_{nj})$ όπου v_{ij} είναι η τιμή της ιδιότητας A_j με $j \in \{1,2, \dots A\}$ όπου A το συνολικό πλήθος των ιδιοτήτων του συνόλου δεδομένων. Μια ιδιότητα μπορεί είναι διακριτή (discrete), δηλαδή να λαμβάνει τιμές από ένα πεπερασμένο σύνολο τιμών, ή συνεχής (continuous) λαμβάνοντας τιμές από ένα άπειρο σύνολο. Ένα παράδειγμα e_j, είναι ένα στιγμιότυπο με προσδιορισμένη κατηγορία. Με άλλα λόγια είναι μια πλειάδα $(v_{1j}, v_{2j}, \dots, v_{nj}, c_j)$, όπου $c_j \in \{c_1, \dots, c_C\}$ η κατηγορία της συγκεκριμένης πλειάδας με C το συνολικό πλήθος των κατηγοριών. Ένα σύνολο εκπαίδευσης (training dataset) είναι ένα σύνολο παραδειγμάτων, συνήθως οργανωμένων σε μορφή πίνακα, με τις στήλες να αντιστοιχούν στις ιδιότητες και τις γραμμές στα παραδείγματα.

2.2.2 Αναπαράσταση κανόνων

Έχοντας ένα σύνολο εκπαίδευσης, ένας αλγόριθμος επαγωγής κανόνων κατασκευάζει κανόνες στην ακόλουθη μορφή:

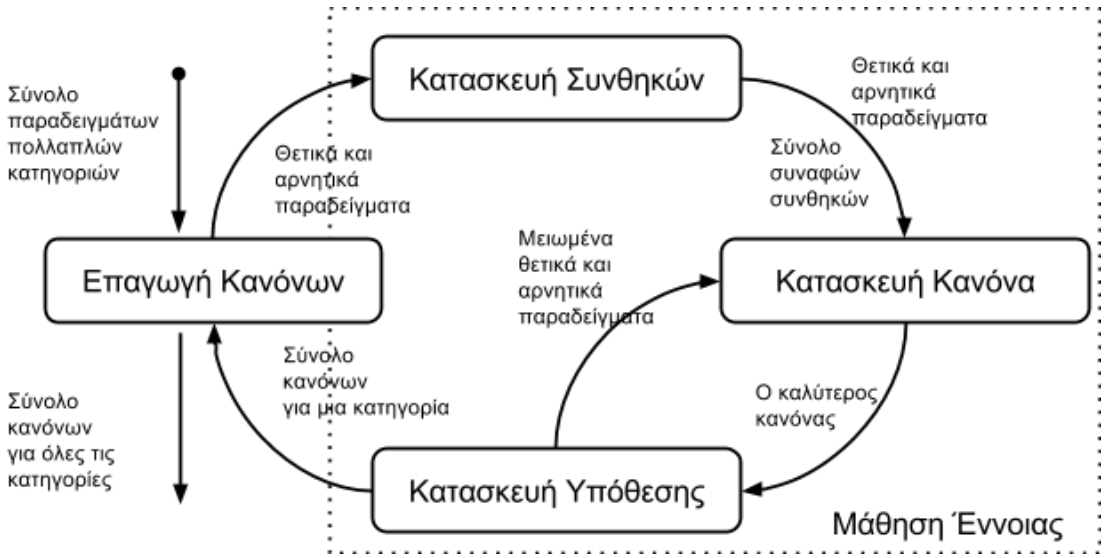
$$AN\ \sigma_1\ KAI\ \ \sigma_2\ KAI\ \dots .KAI\ \sigma_L\ TOTE\ κατηγορία$$

Το τμήμα του κανόνα που οριοθετείται από τον όρο AN, είναι γνωστό ως η προϋπόθεση (antecedent) ή σώμα (body) και περιέχει μια λογική σύζευξη συνθηκών. Το τμήμα που οριοθετείται

από τον όρο TOTE, είναι γνωστό ως συνέπεια (consequent) ή κεφαλή (head) και ορίζει την κλάση που προβλέπει ο συγκεκριμένος κανόνας για στιγμιότυπα που ικανοποιούν την προϋπόθεσή του. Κάθε συνθήκη σ_i είναι μια τριάδα <ιδιότητα, τελεστής, τιμή> και είναι συνήθως συζευγμένη με άλλες συνθήκες με το λογικό τελεστή σύζευξης ΚΑΙ. Το πλήθος L αυτών των συνθηκών καλείται μέγεθος του κανόνα. Ένα παράδειγμα καλύπτεται (covered) από έναν κανόνα αν όλες οι συνθήκες του είναι ίδιες με αυτές του κανόνα. Επίσης ένα παράδειγμα καλύπτεται σωστά (correctly covered) αν καλύπτεται και επιπλέον, η κατηγορία του κανόνα ταυτίζεται με αυτή του παραδείγματος.

2.2.3 Η διαδικασία της επαγωγής

Στην Εικόνα 2.1 φαίνονται τα βήματα της διαδικασίας επαγωγής κανόνων για κατηγοριοποίηση. Το πρόβλημα αυτό ουσιαστικά μετατρέπεται σε μια σειρά από διαδικασίες μάθησης εννοιών (concept learning tasks) ή απλούστερα σε μια σειρά διαδικασιών μάθησης μιας κατηγορίας. Ο αλγόριθμος μάθησης εννοιών αποτελείται από τρεις διαδικασίες που εκτελούνται επαναληπτικά μέχρι να καλυφθεί πλήρως το σύνολο των παραδειγμάτων ή να ικανοποιηθούν άλλα κριτήρια. Αυτές είναι α) η κατασκευή των συνθηκών, β) η κατασκευή των κανόνων και γ) η κατασκευή της υπόθεσης, δηλαδή της λίστας των κανόνων. Για κάθε διαδικασία μάθησης έννοιας υπάρχει ένα σύνολο παραδειγμάτων των οποίων η κατηγορία είναι η έννοια που μαθαίνεται και ένα σύνολο αρνητικών παραδειγμάτων των οποίων η κατηγορία είναι διαφορετική της έννοιας. Το σύνολο των συνθηκών που είναι σχετικές με κάθε έννοια μπορεί να δημιουργηθεί κατά τη διαδικασία της κατασκευής συνθηκών. Στη συνέχεια η διαδικασία της κατασκευής κανόνα χρησιμοποιεί αυτές τις συνθήκες για να κατασκευάσει το σώμα ενός κανόνα για μια συγκεκριμένη κατηγορία.



Εικόνα 2.1 - Η διαδικασία της επαγωγής κανόνων κατηγοριοποίησης

Σε κάθε επανάληψη της κατασκευής της υπόθεσης, το σύνολο των παραδειγμάτων μειώνεται αφού αφαιρούνται τα παραδείγματα που καλύφθηκαν από τον τρέχων κανόνα, ή το τρέχον σύνολο κανόνων που κατασκευάστηκαν στις προηγούμενες επαναλήψεις. Όταν καλυφθούν όλα τα θετικά παραδείγματα, ή έχει ικανοποιηθεί κάποια άλλη συνθήκη τερματισμού, η διαδικασία μάθησης έννοιας τερματίζει. Το σύνολο των κανόνων που περιγράφουν την έννοια συμπεριλαμβάνεται στο τελικό σύνολο κανόνων κατηγοριοποίησης.

Η κατασκευή των κανόνων από τις συνθήκες που έχουν κατασκευαστεί και επιλεχθεί από το προηγούμενο επίπεδο, περιλαμβάνει το σχηματισμό ενός κανόνα από το σύνολο όλων των πιθανών συνδυασμών των μη αλληλοσυγκρουόμενων συνθηκών. Αλληλοσυγκρουόμενες είναι οι συνθήκες των ίδιων ιδιοτήτων, όπως για παράδειγμα στην περίπτωση που η συνθήκη "φύλο=αρσενικό" είναι στον ίδιο κανόνα με τη συνθήκη "φύλο=θηλυκό". Ο κανόνας που προκύπτει από αυτό το επίπεδο, καλύπτει ένα υποσύνολο του χώρου των παραδειγμάτων. Περισσότερα για τη διαδικασία αυτή αναφέρονται στην παράγραφο 2.2.4.

Τέλος, κατά την κατασκευή της υπόθεσης οι κανόνες ομαδοποιούνται ατομικά και σειριακά, χρησιμοποιώντας κάποιον αλγόριθμο όπως τον αλγόριθμο κάλυψης. Με έναν τέτοιο αλγόριθμο, είναι δυνατόν να δημιουργηθούν λίστες κανόνων (decision lists) ή μη διατεταγμένα σύνολα κανόνων (unordered rule sets).

Οι διάφορες προσεγγίσεις για την κατασκευή κανόνων κατηγοριοποίησης που έχουν προταθεί στη βιβλιογραφία, ουσιαστικά προσδιορίζουν τη στρατηγική σύμφωνα με την οποία γίνεται κάθε επιλογή στα προαναφερθέντα επίπεδα. Κάθε στρατηγική περιλαμβάνει την αξιολόγηση των συνθηκών, των κανόνων και των μοντέλων ενάντια σε ολόκληρο ή ένα υποσύνολο των παραδειγμάτων έτσι ώστε να προσδιοριστεί η ποιότητά τους. Η κατασκευή κανόνων κατηγοριοποίησης κατά συνέπεια μπορεί να οριστεί επίσης και ως ένα πρόβλημα αναζήτησης στους χώρους των συνθηκών, των κανόνων και των παραδειγμάτων. Ο χώρος των συνθηκών μπορεί εύκολα να απαριθμηθεί όταν όλες οι ιδιότητες είναι διακριτές και χρησιμοποιείται ένα είδος τελεστή: είναι το πλήθος όλων των τιμών των ιδιοτήτων. Αντίθετα, στην περίπτωση των συνεχών μεταβλητών είναι άπειρο αλλά πρακτικά περιορίζεται στις τιμές που συναντώνται στο σύνολο των παραδειγμάτων. Ο χώρος των κανόνων, όμοια μπορεί να απαριθμηθεί στην περίπτωση των διακριτών ιδιοτήτων και αποτελείται από το συνδυασμό όλων των έγκυρων συνθηκών. Τέλος, το μέγεθος του χώρου των παραδειγμάτων είναι συνήθως γνωστό εκ των προτέρων, ωστόσο κάτι τέτοιο μπορεί να μην ισχύει στα σύνολα δεδομένων μεγάλης κλίμακας, όπου το πλήθος των παραδειγμάτων μπορεί να μην είναι γνωστό εκ των προτέρων.

2.2.4 Η διαδικασία κατασκευής ενός κανόνα

Δεδομένου ενός συνόλου συνθηκών, η διαδικασία κατασκευής ενός κανόνα για μια κατηγορία ανάγεται στην επιλογή του κατάλληλου συνδυασμού των συνθηκών που περιγράφει καλύτερα τη

λογική συσχέτιση μεταξύ των συνθηκών και της κατηγορίας, αν αυτή υπάρχει. Η διαδικασία κατασκευής ενός κανόνα μπορεί να αναχθεί σε ένα πρόβλημα αναζήτησης [34] αφού πρώτα έχει οριστεί α) ο κατάλληλος χώρος αναζήτησης, β) μια στρατηγική αναζήτησης και γ) μια συνάρτηση ποιότητας που αποτιμά τους κανόνες έτσι ώστε να προσδιορίσει πόσο κοντά στη λύση είναι ένας υποψήφιος κανόνας.

Αλγόριθμος 2.1 - Ένας γενικός αλγόριθμος για την κατασκευή ενός κανόνα

```
ΕΙΣΟΔΟΣ:      Ε, ένα σύνολο παραδειγμάτων
ΈΞΟΔΟΣ:       r_best, ο καλύτερος κανόνας για τα συγκεκριμένα παραδείγματα

1. r_best = αρχικοποίησε_κανόνα(Ε)
2. h_best = αξιολόγησε_κανόνα(r)
3. R = {r_best}
4. ΟΣΟ R ≠ Ø
5.   R = επέλεξε_υποψήφιους(R,E)
6.   ΓΙΑ ΚΑΘΕ κανόνα r ∈ R
7.     R' = βελτίωσε_κανόνα(r, E)
8.     ΓΙΑ ΚΑΘΕ κανόνα r'∈ R'
9.       ΑΝ κριτήριο_τερματισμού(r',E) ΤΟΤΕ επόμενος κανόνας
10.      h = αξιολόγησε_κανόνα(r')
11.      R = ταξινόμησε(r', R)
12.      ΑΝ h > h_best ΤΟΤΕ h_best= h, r_best= r'
13.    ΤΕΛΟΣ ΕΠΑΝΑΛΗΨΗΣ
14.  ΤΕΛΟΣ ΕΠΑΝΑΛΗΨΗΣ
15.  R = φιλτράρισε_κανόνες(R, E)
16. ΤΕΛΟΣ ΕΠΑΝΑΛΗΨΗΣ
```

Στον Αλγόριθμος 2.1 περιγράφεται η διαδικασία αναζήτησης στον χώρο των κανόνων για εκείνον τον κανόνα που βελτιστοποιεί το κριτήριο ποιότητας όπως ορίζεται στη συνάρτηση *αξιολόγησε_κανόνα()*. Η ποιότητα του κανόνα εξαρτάται από το πλήθος των θετικών και των αρνητικών παραδειγμάτων που καλύπτονται από τον κανόνα. Ο αλγόριθμος διατηρεί το σύνολο R, που είναι μια λίστα υποψήφιων κανόνων που αρχικοποιείται με τη συνάρτηση *αρχικοποίησε_κανόνα()*. Οι κανόνες στο σύνολο R είναι ταξινομημένοι κατά φθίνουσα σειρά ως προς την ποιότητά τους. Σε κάθε εξωτερική επανάληψη γίνεται η επιλογή των υποψήφιων κανόνων στη συνάρτηση *επέλεξε_υποψήφιους()*, οι οποίοι βελτιώνονται με τη συνάρτηση *βελτιώσε_κανόνα()*. Η βελτίωση ενός κανόνα ουσιαστικά αντιστοιχεί στην αντικατάστασή του από διάδοχους κανόνες, που θα διαφέρουν κατά μια συνθήκη. Κάθε βελτιωμένος κανόνας αξιολογείται και τοποθετείται στη λίστα R, εκτός αν το αποτρέψει η συνάρτηση *κριτήριο_τερματισμού()*. Αν ο νέος κανόνας είναι καλύτερος από τον προηγούμενο καλύτερο τότε ο πρώτος γίνεται ο νέος καλύτερος κανόνας. Η συνάρτηση *φιλτράρισε_κανόνες()*, αφαιρεί αχρείαστους κανόνες για την επόμενη επανάληψη.

Οι ορισμοί των διαφόρων συναρτήσεων στον Αλγόριθμος 2.1 ουσιαστικά καθορίζουν τις προσεγγίσεις που ακολουθούν οι διάφορες λύσεις που έχουν προταθεί στη βιβλιογραφία. Οι συναρτήσεις *αρχικοποίησε_κανόνα* και *βελτίωσε_κανόνα*, ορίζουν τη στρατηγική αναζήτησης, οι συναρτήσεις *επέλεξε_υποψήφιους* και *φιλτράρισε_κανόνες* ορίζουν τον αλγόριθμο αναζήτησης, η *αξιολόγησε_κανόνα* ορίζει την ευρετική συνάρτηση και η *κριτήριο_τερματισμού* ορίζει τον τρόπο αντιμετώπισης του φαινομένου της υπερπροσαρμογής (overfitting), δηλαδή της περιγραφής μιας ακραίας τιμής, ενός λάθους ή θορύβου, αντί της υποκείμενης σχέσης που θα έπρεπε να περιγράφει ένα μοντέλο.

Για την δημιουργία ενός κανόνα, οι περισσότεροι αλγόριθμοι μάθησης χρησιμοποιούν μια από τις ακόλουθες στρατηγικές:

- Εξειδίκευσης (general to specific): όπου οι αλγόριθμοι ξεκινάνε από τον πιο γενικό κανόνα και επαναληπτικά τον εξειδικεύουν όσο οι κανόνες που βρέθηκαν καλύπτουν αρνητικά παραδείγματα. Η εξειδίκευση σταματά όταν ο κανόνας είναι συνεπής. Κατά την αναζήτηση, οι αλγόριθμοι εξειδίκευσης διασφαλίζουν ότι οι υπό εξέταση κανόνες καλύπτουν τουλάχιστον ένα θετικό παράδειγμα. Επίσης, επειδή χρησιμοποιούνται ευρετικοί μηχανισμοί, η συγκεκριμένη στρατηγική ενδείκνυται σε δεδομένα εκπαίδευσης όπου υπάρχει θόρυβος.
- Γενίκευσης (specific to general): που ξεκινάνε από τον πιο εξειδικευμένο κανόνα, και τον γενικεύουν μέχρι να μην μπορεί να γενικευτεί χωρίς να καλύπτει αρνητικά παραδείγματα. Η προσέγγιση αυτή φαίνεται να ταιριάζει στις περιπτώσεις όπου υπάρχουν διαθέσιμα λιγότερα παραδείγματα. Ωστόσο οι αλγόριθμοι που χρησιμοποιούν αυτή τη στρατηγική, επηρεάζονται αρκετά από το θόρυβο στα δεδομένα.

Οι επιλογές για την αναζήτηση του χώρου κανόνων συμπεριλαμβάνουν α) τους αλγορίθμους που χρησιμοποιούν ευρετικούς μηχανισμούς, β) τους εξαντλητικούς αλγόριθμους και γ) τους στοχαστικούς αλγόριθμους:

- Ο πιο συχνά χρησιμοποιούμενος αλγόριθμος αναζήτησης κατά την κατασκευή ενός κανόνα είναι η αναρρίχηση λόφων (hill-climbing). Ο αλγόριθμος αυτός προσπαθεί να βρει τον κανόνα με τη βέλτιστη ποιότητα, προσπαθώντας σε κάθε βήματα κατασκευής του να προσθέσει συνθήκες που θα αυξήσουν την ποιότητά του και σταματώντας όταν δεν είναι δυνατή κάποια περεταίρω βελτίωση. Με άλλα λόγια η αναρρίχηση λόφων προσπαθεί να ανακαλύψει το καθολικό βέλτιστο πραγματοποιώντας τοπικές βελτιώσεις. Το κύριο μειονέκτημα της προσέγγισης αυτής ωστόσο είναι η το φαινόμενο που αναφέρεται στη βιβλιογραφία ως «μυωπία» [35]. Σύμφωνα με το φαινόμενο αυτό, απορρίπτονται όλες οι βελτιώσεις εκτός από την καλύτερη. Έτσι λιγότερο ποιοτικές

επιλογές που όμως ενδεχομένως να οδηγούν στο μέλλον σε πολύ ανώτερους κανόνες αγνοούνται. Για την αντιμετώπιση του προβλήματος αυτού έχει προταθεί η αναζήτηση δέσμης (beam search) [36] [37] η οποία μαζί με τον καλύτερο κανόνα θυμάται και ένα πεπερασμένο πλήθος εναλλακτικών κανόνων μέσα σε μια δέσμη (beam) τους οποίους και βελτιώνει σε κάθε βήμα. Προφανώς αν το πλήθος εναλλακτικών είναι η μονάδα, ο αλγόριθμος μετατρέπεται σε αναρρίχηση λόφων. Η αναζήτηση δέσμης διατηρεί την αποδοτικότητα της αναρρίχησης λόφων, αλλά εξάγει καλύτερα αποτελέσματα αφού εξερευνάει μεγαλύτερο μέρος του χώρου αναζήτησης.

- Οι προσεγγίσεις που χρησιμοποιούν εξαντλητικούς αλγόριθμους είναι εγγυημένο ότι θα βρουν τον κανόνα με την καλύτερη ποιότητα, σε αντίθεση με τους αντίστοιχους ευρετικούς. Ωστόσο είναι πιο ευαίσθητοι στο φαινόμενο της υπερπροσαρμογής κάτι που πρέπει να λαμβάνεται υπόψη κατά την υλοποίησή τους. Υπάρχουν δυο κύριες προσεγγίσεις σε αυτήν την κατηγορία: η αναζήτηση πρώτα-στο-καλύτερο (best-first search) που ουσιαστικά αποτελεί μια αναζήτηση δέσμης με άπειρες δέσμες και η αναζήτηση επιπέδου (level-wise search) που πραγματοποιεί μια κατά πλάτος αναζήτηση βρίσκοντας σε κάθε επανάληψη τον καλύτερο από τους κανόνες με μέγεθος κατά ένα μεγαλύτερο από αυτούς της προηγούμενης επανάληψης.
- Άλλη μια προσέγγιση για την αποφυγή της παραμονής σε τοπικά βέλτιστα είναι η χρήση της στοχαστικής αναζήτησης (stochastic search) που κάνει χρήση της τύχης κατά τη διαδικασία της βελτίωσης του τρέχοντος κανόνα. Κάτι τέτοιο επιτρέπει στον αλγόριθμο να εξερευνήσει εντελώς νέες περιοχές του χώρου αναζήτησης. Στην απλούστερη περίπτωση, η συνάρτηση *βελτίωσε_κανόνα()* θα επιστρέψει έναν τυχαίο κανόνα του χώρου αναζήτησης, ωστόσο υπάρχουν εκδοχές που δίνουν περισσότερες πιθανότητες επιλογής στους κανόνες με καλύτερη ποιότητα.

Ένα πολύ σημαντικό θέμα στη διαδικασία μάθησης είναι η αξιολόγηση της ποιότητας κάθε συνθήκης, έτσι ώστε η διαδικασία της αναζήτησης να οδηγηθεί προς την ανακάλυψη κανόνων που καλύπτουν σωστά όσο το δυνατόν περισσότερα παραδείγματα, ενώ ταυτόχρονα ελαχιστοποιούν τα παραδείγματα που είναι λάθος καλυμμένα. Έχουν προταθεί διάφορες προσεγγίσεις στη βιβλιογραφία μέσα από τη χρήση ευρετικών συναρτήσεων. Οι περισσότερες από αυτές εστιάζονται στην εύρεση μιας μεθόδου επαγωγής που έχει την καλύτερη επίδοση στο μεγαλύτερο εύρος συνόλων δεδομένων, ανεξάρτητα από το είδος των δεδομένων που αναλύονται. Παρόλο που τα περισσότερα από αυτά βασίζονται στη θεωρία πληροφορίας, τη στατιστική ή άλλα σχετικά πεδία, οι σχέσεις μεταξύ τους δεν είναι πλήρως ορισμένες και κατά συνέπεια δεν είναι εύκολο να συμπεράνει κανείς ότι μια συγκεκριμένη ευρετική συνάρτηση είναι καλύτερη από όλες τις υπόλοιπες. Μια αναλυτική πειραματική σύγκριση των διαφορετικών ευρετικών συναρτήσεων που χρησιμοποιούνται για την

κατασκευή κανόνων, μπορεί να βρεθεί στο [38]. Ένα από τα πιο σημαντικά μέτρα για την αποτίμηση της ποιότητας ενός κανόνα είναι η ακρίβεια (precision) που ορίζεται ως:

$$precision(r) = \frac{TP}{P}$$

όπου TP είναι το συνολικό πλήθος των σωστά καλυμμένων παραδειγμάτων και το P είναι το συνολικό πλήθος καλυμμένων παραδειγμάτων. Το πρόβλημα με την ακρίβεια είναι ότι δεν είναι δηλωτική του μεγέθους των καλυμμένων παραδειγμάτων. Για παράδειγμα έστω δυο κανόνες. Ο κανόνας Α έχει δυο σωστά καλυμμένα παραδείγματα και ο κανόνας Β έχει εκατό σωστά καλυμμένα παραδείγματα. Και στις δυο περιπτώσεις, το μέτρο της ακρίβειας θα δώσει τη μέγιστη τιμή του, ωστόσο δεν θα αντιπροσωπεύει σωστά το ποσοστό κάλυψης ολόκληρου του συνόλου δεδομένων. Στο πλαίσιο των δεδομένων μεγάλης κλίμακας, η αξιολόγηση της ποιότητας των κανόνων πρέπει να γίνεται σε ολόκληρο το σύνολο παραδειγμάτων έτσι ώστε να κατασκευαστούν κανόνες ποιότητας σε όσο το δυνατόν λιγότερες επαναλήψεις γίνεται. Κάτω από αυτό το πρίσμα, η ακρίβεια ως μέτρο αξιολόγισης δεν είναι αξιόπιστη.

Στο προηγούμενο παράδειγμα, αν τόσο τα σωστά όσο και τα λανθασμένα καλυμμένα παραδείγματα είναι λίγα, η πρόσθεση ενός εικονικού παραδείγματος μπορεί να αλλάξει την τιμή της αξιολόγησης σημαντικά, αλλά κάτι τέτοιο δεν θα ισχύει αν ένα από αυτά (ή και τα δυο) είναι υψηλά. Στην πραγματικότητα αυτή είναι η σκέψη πίσω από το μέτρο Laplace:

$$Laplace(r) = \frac{TP + 1}{P + 2}$$

Το μέτρο ασυμπτωτικά προσεγγίζει το ένα αν το πλήθος των σωστά καλυμμένων παραδειγμάτων αυξηθεί, αλλά με πεπερασμένο πλήθος δεδομένων δεν θα φτάσει ποτέ το ένα. Ένα άλλο μέτρο που χρησιμοποιείται συχνά σε διάφορες προσεγγίσεις όπως η ID3 [39] και η CN2 [37], χρησιμοποιεί το λογάριθμο του πλήθους των παραδειγμάτων που έχει καλύψει ορθά ο κανόνας προς το πλήθος των καλυμμένων παραδειγμάτων, που μετρά το περιεχόμενο της πληροφορίας κατά την ανάθεση της συνέπειας από τα καλυμμένα παραδείγματα:

$$Info(r) = -\log_2 \frac{TP}{P}$$

Αυτή η ποσότητα μπορεί να χρησιμοποιηθεί σε μια πιο πολύπλοκη σχέση που είναι γνωστή ως εντροπία (entropy):

$$Entropy(r) = -\sum_{i=1}^{C} \frac{TP}{P} \log_2 \frac{TP}{P}$$

Όπου C είναι το συνολικό πλήθος των κατηγοριών ενώ τα TP και P έχουν οριστεί παραπάνω.

Ένα μέτρο αξιολόγησης της ποιότητας ενός κανόνα που λαμβάνει υπόψη την κάλυψη των παραδειγμάτων του μπορεί να συνδυάσει την κανονικοποιημένη εντροπία σε συνδυασμό με την κάλυψη ως εξής:

$$\frac{Entropy(C) - Entropy(r)}{Entropy(C)} \times \frac{P}{E}$$

Όπου C το πλήθος των κατηγοριών, r ο κανόνας, P το πλήθος των καλυμμένων παραδειγμάτων και E το συνολικό πλήθος των παραδειγμάτων. Άλλα γνωστά μέτρα που αξιολογούν την ακρίβεια ενός κανόνα και ταυτόχρονα λαμβάνουν υπόψη τους την κάλυψη είναι η εκτίμηση-m [40], το μέτρο-g [40], το κέρδος, το Cohen [41], το C2 [41], η ειδικότητα και η ευαισθησία και η βεβαρημένη σχετική ακρίβεια [40].

2.3 Το πρόβλημα της διακριτοποίησης δεδομένων

Πολλοί αλγόριθμοι μηχανικής μάθησης και εξόρυξης γνώσης από δεδομένα, όπως αυτοί της επαγωγής κανόνων κατηγοριοποίησης, της επαγωγής κανόνων συσχέτισης και των δικτύων Bayes μπορούν να χειριστούν μόνο διακριτές ιδιότητες. Ωστόσο πάρα πολλές εφαρμογές περιλαμβάνουν ή αποτελούνται εξολοκλήρου από δεδομένα με συνεχείς ιδιότητες. Κατά συνέπεια πριν την εκτέλεσή τέτοιων αλγορίθμων σε προβλήματα με αριθμητικά δεδομένα, είναι αναγκαία η κωδικοποίηση κάθε συνεχούς ιδιότητας σε διακριτή. Η διαδικασία αυτή είναι γνωστή ως διακριτοποίηση (discretization) και είναι απαραίτητη κατά την προεπεξεργασία των δεδομένων προς ανάλυση, όχι μόνο γιατί ορισμένες τεχνικές μάθησης δεν μπορούν να χειριστούν συνεχείς ιδιότητες, αλλά γιατί α) τα δεδομένα που αναπαρίστανται σε ένα σύνολο διαστημάτων είναι πιο εύκολα ερμηνεύσιμα από τους ανθρώπους [42], β) το μέγεθος των δεδομένων μειώνεται, αφού τιμές από ένα τεράστιο (θεωρητικά άπειρο) αριθμητικό πεδίο αντιστοιχίζονται σε ένα κατά πολύ μικρότερο σύνολο διαστημάτων και κατά συνέπεια η διαδικασία της μάθησης επιταχύνεται, γ) τα διαστήματα που προκύπτουν μπορούν να αναδείξουν μη γραμμικές συσχετίσεις και δ) οποιοσδήποτε θόρυβος υπάρχει στα δεδομένα μειώνεται. Ωστόσο οποιαδήποτε μέθοδος διακριτοποίησης οδηγεί κατά γενική ομολογία σε απώλεια πληροφορίας, θέτοντας έτσι σαν στόχο για κάθε προσέγγιση, την ελαχιστοποίησή της.

Η πιο αποδοτική διακριτοποίηση είναι συχνά αυτή που προκύπτει από έναν ειδικό στον τομέα του προβλήματος, αφού αυτός θα είναι σε θέση να διακρίνει τα βέλτιστα όρια των διαστημάτων και να τα προσαρμόσει ανάλογα με τις απαιτήσεις του προβλήματος. Ωστόσο αυτή η προσέγγιση δεν είναι εφικτή στην πλειονότητα των προβλημάτων μηχανικής μάθησης, επειδή α) μπορεί να μην υπάρχουν ειδικοί, β) δεν υπάρχει πρότερη γνώση πάνω στο αντικείμενο, ή γ) το μέγεθος των δεδομένων είναι απαγορευτικό για να αναλυθεί από έναν ειδικό και κατά συνέπεια το κόστος θα αυξανόταν δραματικά. Για τους λόγους αυτούς, έχουν προταθεί και αναπτυχθεί αυτόματες μέθοδοι

διακριτοποίησης ιδιοτήτων για την εύρεση των βέλτιστων διαστημάτων για ένα συγκεκριμένο πρόβλημα.

2.3.1 Ορισμός

Σύμφωνα με την εργασία [43] το πρόβλημα της διακριτοποίησης ορίζεται ως εξής: με είσοδο ένα σύνολο δεδομένων αποτελούμενο από N παραδείγματα και C κατηγορίες, ένας αλγόριθμος διακριτοποίησης στοχεύει στη διαίρεση μιας ιδιότητας A αυτού του συνόλου, σε m διακριτά διαστήματα $D = \{[d_0, d_1], (d_1,d_2], (d_2,d_3], \dots, (d_{m-1},d_m] \}$, όπου d_0είναι η ελάχιστη τιμή, d_mείναι η μέγιστη τιμή και $d_i < d_{i+1}$, για $i = 0,1, \dots m-1$. Ένα τέτοιο αποτέλεσμα D καλείται σχήμα διακριτοποίησης (discretization schema) πάνω στην ιδιότητα A και $P = \{d_1, d_2, \dots, d_{m-1}\}$ είναι το σύνολο των οριακών σημείων (cut points) της ιδιότητας A.

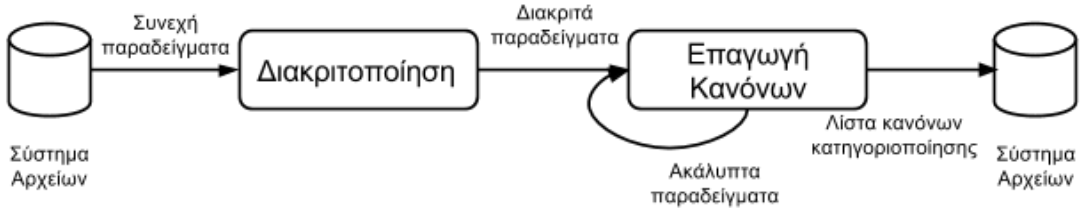
Στη βιβλιογραφία είναι διαθέσιμο ένα μεγάλο πλήθος μεθόδων διακριτοποίησης. Η επιλογή της κατάλληλης μεθόδου διακριτοποίησης, παίζει μεγάλο ρόλο στην ακρίβεια και την απλότητα του μοντέλου εκπαίδευσης που θα προκύψει από τον αλγόριθμο μάθησης. Οι διάφορες μέθοδοι που έχουν προταθεί μπορούν να κατηγοριοποιηθούν λαμβάνοντας υπόψη πολλαπλά κριτήρια, όπως τον ευρετικό μηχανισμό που χρησιμοποιούν, αν χρησιμοποιούν όλες τις ιδιότητες ή μόνο μία, αν είναι με επίβλεψη ή χωρίς, αν είναι καθολικές ή τοπικές, αν είναι στατικές ή δυναμικές κ.α.. Όλα αυτά τα κριτήρια αποτελούν τη βάση των ταξινομήσεων που έχουν προταθεί μέχρι τώρα στη βιβλιογραφία οι οποίες μπορούν να βρεθούν σε βιβλιογραφικές επισκοπήσεις όπως τις [43], [44], [45]και [46].

Ένας αλγόριθμος διακριτοποίησης μπορεί να είναι στατικός (static) [47] ή δυναμικός (dynamic) [48] ανάλογα με το αν ενεργεί στα πλαίσια ενός αλγόριθμου κατηγοριοποίησης ή αν ενεργεί ανεξάρτητα. Ένας στατικός αλγόριθμος διακριτοποίησης εκτελείται πριν τη διαδικασία κατηγοριοποίησης, ενώ ένας αντίστοιχος δυναμικός ενεργεί κατά τη διάρκεια κατασκευής του μοντέλου κατηγοριοποίησης. Ένας αλγόριθμος διακριτοποίησης επίσης μπορεί να είναι μονομεταβλητός (univariate) [49] αν επεξεργάζεται κάθε ιδιότητα ενός συνόλου δεδομένων ανεξάρτητα από τις υπόλοιπες ή πολυμεταβλητός (multivariate) [50] αν λαμβάνει υπόψη όλες τις ιδιότητες κατά τη διαδικασία της διακριτοποίησης. Ανάλογα με το αν λαμβάνει υπόψη την κατηγορία των παραδειγμάτων, ένας αλγόριθμος διακριτοποίησης μπορεί να χαρακτηρίζεται ως αλγόριθμος με επίβλεψη (supervised), ή χωρίς επίβλεψη (unsupervised) [51]. Επίσης η δημιουργία των τελικών διαστημάτων μπορεί να γίνεται με το διαδοχικό διαχωρισμό (splitting) μεγάλων διαστημάτων σε μικρότερα ή τη συγχώνευση (merging) μικρών διαστημάτων σε μεγαλύτερα. Αν ένας αλγόριθμος προβαίνει στη δημιουργία ενός διαστήματος λαμβάνοντας υπόψη όλα τα δεδομένα του συνόλου καλείται καθολικός (global) ενώ αν χρειάζεται μεμονωμένα υποσύνολα από τα δεδομένα καλείται τοπικός (local) [48]. Τέλος άλλο ένα σημαντικό χαρακτηριστικό των αλγόριθμων διακριτοποίησης είναι το μέτρο αξιολόγησης δυο υποψήφιων σχημάτων. Οι πιο συχνά χρησιμοποιούμενες οικογένειες μέτρων είναι αυτές που βασίζονται στη θεωρία πληροφορίας

(information theory) [48], τη στατιστική (statistics) [52] και τα προσεγγιστικά σύνολα (rough sets) [53].

2.4 Επαγωγή κανόνων κατηγοριοποίησης με το μοντέλο Απεικόνισης/Μείωσης

Στην παράγραφο αυτή παρουσιάζεται η προτεινόμενη προσέγγιση στο πρόβλημα επαγωγής κανόνων κατηγοριοποίησης. Μια επισκόπηση της προσέγγισης αυτής από υψηλό επίπεδο, φαίνεται στην Εικόνα 2.2. Η προσέγγιση βασίζεται στο μοντέλο Απεικόνισης/Μείωσης αποτελείται από δύο βήματα, το πρώτο, που περιγράφεται στην παράγραφο 2.4.1, αναλαμβάνει τη διακριτοποίηση όσων από τα χαρακτηριστικά του συνόλου δεδομένων εισόδου είναι συνεχή. Η διαδικασία αυτή είναι ένα μεμονωμένο έργο Απεικόνισης/Μείωσης. Αντίθετα το δεύτερο βήμα αναλαμβάνει την κατασκευή της λίστας κανόνων κατηγοριοποίησης και είναι μια επαναληπτική διαδικασία που περιγράφεται στην παράγραφο 2.4.2.



Εικόνα 2.2 - Μια επισκόπηση της προτεινόμενης προσέγγισης επαγωγής κανόνων κατηγοριοποίησης

2.4.1 Η διαδικασία διακριτοποίησης

Η προτεινόμενη προσέγγιση διακριτοποίησης είναι μια στατική, μονομεταβλητή και καθολική προσέγγιση με επίβλεψη, που συγχωνεύει αρχικά μικρά διαστήματα σε μεγαλύτερα τελικά. Η προσέγγιση βασίζεται στην τεχνική Απεικόνισης/Μείωσης και στοχεύει σε δεδομένα μεγάλης κλίμακας. Δεδομένου ενός συνόλου παραδειγμάτων σε αριθμητική μορφή $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ όπου $x_i \in R^d$ με d το πλήθος των ιδιοτήτων και $y_i \in \{1, \dots, c\}$, όπου c το πλήθος των κατηγοριών, ο στόχος είναι η εύρεση ενός συνόλου οριακών σημείων $P_i = \{d_{i,1}, d_{i,2}, \dots, d_{i,m-1}\}$ για κάθε ιδιότητα του συνόλου $i = 0,1, \dots, d$.

Από μια οπτική υψηλού επιπέδου, η προσέγγιση αποτελείται από ένα μόνο έργο Απεικόνισης/Μείωσης κάτι που σημαίνει ότι η όλη διαδικασία της διακριτοποίησης πραγματοποιείται σε μια μοναδική μαζική επεξεργασία (batch mode). Το πλεονέκτημα της προσέγγισης έγκειται στο χαμηλό κόστος μνήμης και επικοινωνίας και την απουσία επαναλήψεων για την επίτευξη της διακριτοποίησης. Η προτεινόμενη προσέγγιση είναι προσαρμοσμένη στις ειδικές απαιτήσεις του μοντέλου προγραμματισμού της Απεικόνισης/Μείωσης ενώ ταυτόχρονα

εκμεταλλεύεται τους εγγενείς μηχανισμούς διαμοιρασμού δεδομένων και ταξινόμησης που συμβαίνουν μεταξύ των φάσεων της Απεικόνισης και της Μείωσης.

Η διαδικασία της Απεικόνισης κατασκευάζει για κάθε ιδιότητα ένα σύνολο από B διαστήματα

$$\{(min_1, p_{1,}, cd_1, max_1), (min_2, p_{2,}, cd_2, max_2), \dots, (min_B, p_B, cd_B, max_B)\}$$

πραγματικών αριθμών, όπου B είναι ένας ακέραιος που δίνεται σαν παράμετρος από το χρήστη, $\min_i$ το κάτω όριο του διαστήματος, $\max_i$ το άνω όριο του διαστήματος, p_i ο σταθμισμένος μέσος του διαστήματος και cd_i η κατανομή κατηγοριών των παραδειγμάτων του διαστήματος. Η χρήση του σταθμισμένου μέσου κατά τον έλεγχο για την ενσωμάτωση μιας νέας τιμής σε κάποιο από τα διαθέσιμα διαστήματα, έχει προταθεί στην εργασία [54]. Το σύνολο διαστημάτων αποτελεί μια συμπιεσμένη αναπαράσταση των δεδομένων που βρίσκονται στο τοπικό κομμάτι δεδομένων που επεξεργάζεται κάθε εργασία Απεικόνισης. Κάθε δεδομένη χρονική στιγμή, το πλήθος των παραδειγμάτων k που αναπαρίστανται από ένα διάστημα δίνεται από το άθροισμα των στοιχείων της κατανομής κατηγοριών cd_i.

Ο ψευδοκώδικας της εργασίας Απεικόνισης φαίνεται στον Αλγόριθμος 2.2. Η συνάρτηση απεικόνισης λαμβάνει ένα παράδειγμα με κάθε κλήση της στη μορφή <κλειδί, τιμή> όπου το κλειδί αντιστοιχεί σε έναν προσδιοριστή (ένα id ή κάποιο άλλο αναγνωριστικό) και η τιμή αντιστοιχεί στο πραγματικό παράδειγμα. Η εργασία της Απεικόνισης διατηρεί έναν πίνακα κατακερματισμού με κλειδιά τις ιδιότητες του συνόλου δεδομένων και τιμές τα διαστήματα των ιδιοτήτων για το συγκεκριμένο κομμάτι εισόδου που λαμβάνει. Για κάθε παράδειγμα που περνάει σαν είσοδος στη συνάρτηση απεικόνισης, δημιουργείται ένα νέο διάστημα για κάθε ιδιότητα, που περιέχει σαν μέγιστη, ελάχιστη και σταθμισμένο μέσο την τιμή της ιδιότητας του παραδείγματος, ενώ η κατανομή κλάσεων θα περιέχει μόνο την αντίστοιχη κατηγορία. Το νέο αυτό διάστημα και η υπάρχουσα λίστα διαστημάτων που αντιστοιχούν σε αυτή την ιδιότητα περνούν στην είσοδο της συνάρτησης *ενημέρωσε_διαστήματα()*. Η συνάρτηση αυτή ουσιαστικά υπολογίζει την απόσταση του σταθμισμένου μέσου του νέου διαστήματος με κάθε σταθμισμένο μέσο των υπαρχόντων διαστημάτων για τη συγκεκριμένη ιδιότητα και συγχωνεύει το νέο διάστημα με το κοντινότερό του. Η συγχώνευση περιλαμβάνει την ενημέρωση του σταθμισμένου μέσου, της κατανομής κατηγοριών και του μέγιστου και του ελάχιστου. Στη συνέχεια τα διαστήματα ταξινομούνται ως προς το σταθμισμένο μέσο τους.

Η εργασία Απεικόνισης ουσιαστικά αποτελεί μια συμπίεση των παραδειγμάτων. Με το πέρας κάθε εργασίας Απεικόνισης αντί να εξαχθεί το σύνολο των τοπικών παραδειγμάτων, εξάγονται B στο πλήθος διαστήματα που περιέχουν όσες πληροφορίες χρειάζονται για να δημιουργηθούν τα τελικά διαστήματα από τις εργασίες Μείωσης.

Αλγόριθμος 2.2 - Η διαδικασία Απεικόνισης της προτεινόμενης μεθόδου διακριτοποίησης

```
Είσοδος:      <k1, παράδειγμα>
              Β: πλήθος διαστημάτων
Έξοδος:       <ιδιότητα, <διάστημα>>
1. {{διάστημα}a }= {}
2. Συνάρτηση απεικόνιση(κ1,παράδειγμα, ιδιότητα, διάστημα)
3.   ΓΙΑ ΚΑΘΕ ιδιότητα a
4.     διάστημαa = (τιμή^a_παράδειγμα, τιμή^a_παράδειγμα, cd_a, τιμή^a_παράδειγμα)
5.     ΑΝ πλήθος διαστημάτων < Β
6.       Πρόσθεσε το διάστημαa στο {{διάστημα}a }
7.     ΑΛΛΙΩΣ
8.       {{διάστημα}a }= ενημέρωσε_διαστήματα({διάστημα}a, διάστημαa)
9.   ΤΕΛΟΣ ΕΠΑΝΑΛΗΨΗΣ
10.
11. Συνάρτηση ολοκλήρωση()
12.   ΓΙΑ ΚΑΘΕ ιδιότητα a
13.    ΓΙΑ ΚΑΘΕ διάστημα στη λίστα {διάστημα}a
14.      εκπομπή(a, διάστημα)
15.    ΤΕΛΟΣ ΕΠΑΝΑΛΗΨΗΣ
16.  ΤΕΛΟΣ ΕΠΑΝΑΛΗΨΗΣ
17.
18. Συνάρτηση ενημέρωσε_διαστήματα({διάστημα}a, δa)
19.   Βρες το κοντινότερο διάστημα δmin στο δa από τη λίστα {διάστημα}a
20.   Αντικατέστησε τα διαστήματα δmin και δa με το
21.                δ(min_δmin, (p_δmin*k_δmin + p_δa*k_δa)/(k_δmin+k_δa), cd_δmin+α, max_δα)
22.   Ταξινόμησε τη λίστα {διάστημα}a
```

Ο τρόπος συγχώνευσης των διαστημάτων δείχνει ανοχή σε ακραίες τιμές, αρκεί φυσικά να έχει επιλεχθεί ένα λογικά μεγάλο πλήθος διαστημάτων. Με την είσοδο ενός παραδείγματος στη συνάρτηση απεικόνισης, απομονώνονται οι τιμές κάθε ιδιότητας και δημιουργούνται αντίστοιχα διαστήματα που περιέχουν μόνο τις τιμές αυτές. Η διαδικασία αυτή συνεχίζεται (γραμμή 5) μέχρι να σχηματιστούν Β στο πλήθος διαστήματα. Στη συνέχεια κάθε νέα τιμή θα πρέπει να συγχωνευτεί με το κοντινότερο σε αυτήν από τα Β διαστήματα (συνάρτηση ενημέρωσε_διαστήματα). Η επιλογή του πλήθους των διαστημάτων είναι προφανώς αντιστρόφως ανάλογη του επιπέδου λεπτομέρειας στη συμπίεση των δεδομένων: όσο πιο πολλά διαστήματα επιλεχθούν, τόσο πιο λεπτομερής θα είναι η περιγραφή του συνόλου δεδομένων, αλλά τόσο περισσότερη μνήμη θα χρειάζεται σε κάθε εργασία Απεικόνισης και τόσα περισσότερα δεδομένα θα πρέπει να μεταφερθούν από τις εργασίες Απεικόνισης, στις εργασίες Μείωσης.

Στη συνέχεια η εργασία Μείωσης λαμβάνει σαν είσοδο το σύνολο όλων των διαστημάτων για μια ιδιότητα από όλες τις εργασίες Απεικόνισης. Σε κάθε συνάρτηση μείωσης πραγματοποιεί δυο

ελέγχους: α) ελέγχει αν δυο διαδοχικά διαστήματα περιέχονται το ένα στο άλλο και αν ναι τα συγχωνεύει και β) ελέγχει αν δυο διαδοχικά διαστήματα έχουν την ίδια πλειοψηφούσα κατηγορία (από τις κατανομές κατηγοριών) οπότε και τα συγχωνεύει. Η συγχώνευση διαδοχικών διαστημάτων με την ίδια κατηγορία συμβαίνει με τον ίδιο τρόπο που γίνεται η δημιουργία διαστημάτων από τη συγχώνευση τιμών που έχει προταθεί στον αλγόριθμο 1R [55]. Η διαφορά με την προσέγγιση αυτή ωστόσο έγκειται στο γεγονός ότι σε αντίθεση με τον 1R, όπου απαιτείται να υπάρχουν τουλάχιστον 6 συνεχόμενα παραδείγματα της ίδιας κλάσης, στην προτεινόμενη προσέγγιση διακριτοποίησης δεν υπάρχει ελάχιστος απαιτούμενος αριθμός στιγμιότυπων σε κάθε διάστημα. Η εργασία Μείωσης για την προτεινόμενη προσέγγιση φαίνεται στον Αλγόριθμος 2.3.

Αλγόριθμος 2.3 - Η εργασία μείωσης της προτεινόμενης μεθόδου διακριτοποίησης

Είσοδος: <ιδιότητα, <διάστημα>>
Έξοδος: <ιδιότητα, <διάστημα>>
1. Συνάρτηση μείωση(<ιδιότητα, <διάστημα>>, <ιδιότητα, <διάστημα>>)
2. ΓΙΑ ΚΑΘΕ διάστημα δ_i στη λίστα <διάστημα>
3. ΑΝ το δ_i περιέχει ή περιέχεται στο δ_{i+1} ή έχουν την ίδια κατηγορία
4. Αντικατέστησε τα διαστήματα δ_i και δ_{i+1} με το
5. $$\delta(min_{\delta i}, \frac{p_{\delta i} * k_{\delta i} + p_{\delta a} * k_{\delta a}}{k_{\delta i} + k_{\delta a}}, cd_{\delta i + \delta i+1}, max_{\delta i+1})$$
6. ΑΛΛΙΩΣ
7. εκπομπή(ιδιότητα, δ_i)

Ένα προαπαιτούμενο για την ορθή λειτουργία του Αλγόριθμος 2.3 είναι η λίστα διαστημάτων <διάστημα> στην τιμή του ζευγαριού κλειδιού-τιμής της εισόδου να είναι ταξινομημένη. Εφόσον η συνάρτηση μείωσης ελέγχει δυο διαδοχικά διαστήματα για να αποφασίσει αν το ένα περιέχεται στο άλλο, η σύγκριση αυτή δεν θα είχε νόημα αν τα διαστήματα δεν ήταν ταξινομημένα. Ωστόσο, προεπιλεγμένα το μοντέλο της Απεικόνισης/Μείωσης δεν ταξινομεί τις τιμές των ζευγαριών στη φάση του Μοιράσματος και της Ταξινόμησης αλλά τα κλειδιά. Για να αποφευχθεί μια ολοκληρωτική φόρτωση των τιμών στη μνήμη της εργασίας Μείωσης, κάτι που μπορεί να έχει σαν συνέπεια την εξάντλησή της, είναι δυνατόν να γίνει χρήση της προεπιλεγμένης συμπεριφοράς του μοντέλου της Απεικόνισης/Μείωσης στη φάση του Μοιράσματος και της Ταξινόμησης. Ενσωματώνοντας ένα κομμάτι της τιμής στο ενδιάμεσο κλειδί, διαμορφώνοντας έτσι ένα σύνθετο κλειδί, το σύστημα εκτέλεσης μαζί με τα κλειδιά θα ταξινομήσει και τις τιμές. Για το λόγο αυτό, το κλειδί <ιδιότητα>, πέραν του προσδιοριστή της ιδιότητας που αποτελεί το φυσικό κλειδί, θα πρέπει να περιέχει και την ελάχιστη και τη μέγιστη τιμή του διαστήματος που μεταδίδεται από την Απεικόνιση στη Μείωση. Επιπρόσθετα, θα πρέπει να οριστεί και η σειρά ταξινόμησης του ενδιάμεσου κλειδιού, έτσι ώστε να ταξινομηθεί πρώτα η ιδιότητα και στη συνέχεια η ελάχιστη και η μέγιστη τιμή του διαστήματος. Τέλος θα πρέπει να οριστεί και ένας Διαχωριστής, έτσι ώστε όλα τα κλειδιά που αφορούν την ίδια ιδιότητα να μοιραστούν στην ίδια εργασία Μείωσης. Έτσι οι τιμές θα φτάσουν στη συνάρτηση

μείωσης ταξινομημένες, έχοντας ταξινομηθεί από το ίδιο το περιβάλλον εκτέλεσης, έναν τομέα όπου το μοντέλο υπερτερεί, αφού αποτελεί κύριο χαρακτηριστικό του [56].

2.4.2 Η διαδικασία επαγωγής κανόνων κατηγοριοποίησης

Η στρατηγική κάλυψης για την επαγωγή κανόνων προέρχεται από την οικογένεια αλγόριθμων AQ [57] και έχει μελετηθεί και επεκταθεί εκτενώς από τότε. Οι κλασικοί αλγόριθμοι κάτω από αυτή τη στρατηγική κατασκευάζουν σύνολα κανόνων επαναληπτικά, με κάθε επανάληψη να βρίσκει έναν κανόνα που καλύπτει ένα τμήμα του συνόλου παραδειγμάτων, να αφαιρεί τα παραδείγματα που καλύπτονται από τον κανόνα και να συνεχίζει με την εκμάθηση ενός άλλου κανόνα που καλύπτει τα εναπομείναντα παραδείγματα. Τα σύνολα των κανόνων που κατασκευάζονται μπορεί να είναι διατεταγμένα ή και όχι. Η διάταξή τους ωστόσο καθορίζει τη διαδικασία με την οποία γίνεται η κατηγοριοποίηση νέων στιγμιότυπων. Στην περίπτωση που η λίστα είναι διατεταγμένη, ο έλεγχος κάλυψης σταματά όταν ο πρώτος που καλύπτει ένα στιγμιότυπο βρεθεί. Αντίθετα, όταν η λίστα είναι μη διατεταγμένη, ο έλεγχος συνεχίζεται και η κατηγορία που ανατίθεται στο στιγμιότυπο είναι η πλειοψηφούσα κατηγορία μεταξύ των κλάσεων που καταδεικνύουν οι κανόνες που καλύπτουν το στιγμιότυπο.

Η προτεινόμενη προσέγγιση, είναι ένας αλγόριθμος ταξινόμησης κανόνων που ακολουθεί την στρατηγική κάλυψης και βασίζεται στην τεχνική Απεικόνισης/Μείωσης για να παράξει ένα μοντέλο από σύνολα δεδομένων μεγάλης κλίμακας. Η λογική πίσω από αυτή την προσέγγιση είναι η ελαχιστοποίηση των επαναλήψεων που πρέπει να πραγματοποιηθούν για να δημιουργηθεί το μοντέλο. Αυτές με τη σειρά τους, μπορούν να μεταφραστούν σε λιγότερα έργα Απεικόνισης/Μείωσης. Κάτι τέτοιο μπορεί να επιτευχθεί, επιλέγοντας τις συνθήκες εκείνες που μεγιστοποιούν την κάλυψη του συνόλου των παραδειγμάτων και ταυτόχρονα των σωστά καλυμμένων παραδειγμάτων. Υποθέτοντας ότι οι κανόνες που αποτελούνται από ένα συνδυασμό των συνθηκών αυτών, κατά πάσα πιθανότητα θα έχουν υψηλή κάλυψη και θα καλύπτουν σωστά περισσότερα παραδείγματα, το σύνολο παραδειγμάτων θα καλυφθεί με περισσότερη ακρίβεια σε λιγότερες επαναλήψεις.

Από ένα υψηλό επίπεδο, η προσέγγιση ακολουθεί τη συνηθισμένη προσέγγιση επαναληπτικών υπολογισμών στην τεχνική Απεικόνισης/Μείωσης, συνδυάζοντας δυο απλά έργα μέσα σε μια κεντρική διεργασία που συγκεντρώνει τα ενδιάμεσα αποτελέσματά τους και ενεργεί πάνω σε αυτά για να παράξει μια διατεταγμένη λίστα κανόνων. Αυτή η κεντρική διεργασία, που είναι γνωστή στη βιβλιογραφία ως οδηγός (driver), ενορχηστρώνει την παραγωγή των κανόνων, αρχικοποιώντας τα έργα Απεικόνισης/Μείωσης και παρέχοντας τις όποιες επιπρόσθετες πληροφορίες χρειάζονται σε αυτές για να ολοκληρώσουν επιτυχώς τις εργασίες Απεικόνισης και Μείωσης. Να σημειωθεί ότι ο αλγόριθμος υποστηρίζει μόνο ονομαστικές ιδιότητες και κατά συνέπεια τα αριθμητικά σύνολα δεδομένων θα πρέπει να υποστούν κάποια διαδικασία διακριτοποίησης σε προγενέστερο στάδιο,

όπως αυτή της προηγούμενης παραγράφου. Ο ψευδοκώδικας για τον οδηγό φαίνεται στον Αλγόριθμος 2.4.

Αναλυτικότερα, ο οδηγός α) λαμβάνει δυο παραμέτρους από το χρήστη: το ποσοστό των παραδειγμάτων που ο χρήστης ανέχεται να παραμείνουν ακάλυπτα και το μέγιστο μέγεθος των κανόνων, β) παράγει μια άδεια λίστα κανόνων, γ) ξεκινά μια επανάληψη που αποτελείται από δυο έργα Απεικόνισης/Μείωσης τα οποία με κάθε επανάληψη παράγουν έναν κανόνα και δ) τερματίζει με μια μόνο εργασία Απεικόνισης/Μείωσης που βρίσκει τον προεπιλεγμένο κανόνα.

Αλγόριθμος 2.4 - Ο οδηγός της προτεινόμενης προσέγγισης

```
Είσοδος:      Ποσοστό ακάλυπτων παραδειγμάτων,
              Μέγιστο πλήθος συνθηκών ανά κανόνα
Έξοδος:       Λίστα κανόνων

1.k ← πλήθος ιδιοτήτων
2.   επανέλαβε
3.      Έργο1: Βρες όλες τις συνθήκες και αξιολόγησέ τες
4.      Επέλεξε τις k καλύτερες συνθήκες
5.      Έργο2: Βρες όλους τους κανόνες με τις k καλύτερες συνθήκες και
   αξιολόγησέ τες
6.      Λίστα κανόνων ←  Επέλεξε τον καλύτερο κανόνα
7.   μέχρι (τα ακάλυπτα στιγμιότυπα <= συνολικά στιγμιότυπα * ποσοστό
   παραδειγμάτων προς κάλυψη)
8. Αν( ακάλυπτα στιγμιότυπα > 0) τότε
9.  Έργο3: Βρες τον προεπιλεγμένο κανόνα
```

Το πρώτο έργο βρίσκει όλες τις συνθήκες που καλύπτονται από το σύνολο παραδειγμάτων και υπολογίζει την κατανομή συχνοτήτων των κλάσεων για όλα τα παραδείγματα που καλύπτει κάθε συνθήκη. Σύμφωνα με αυτές τις κατανομές η πρώτη εργασία αξιολογεί κάθε συνθήκη αθροίζοντας τις συχνότητες όλων των κατηγοριών για τη συγκεκριμένη συνθήκη. Ο οδηγός τότε λαμβάνει τη λίστα των συνθηκών με την αντίστοιχη τιμή αξιολόγησης του καθενός, ταξινομεί τη λίστα ως προς την τιμή αυτή και διαλέγει τις k-καλύτερες συνθήκες, όπου k=A-1 (το πλήθος των ιδιοτήτων μείον ένα). Το k επιλέγεται καθ' αυτόν τον τρόπο, επειδή είναι επιθυμητό να δημιουργηθούν κανόνες χρησιμοποιώντας όλες τις ιδιότητες. Αν όλες οι k συνθήκες είναι της ίδιας ιδιότητας, τότε ο μοναδικός καλύτερος κανόνας μιας συνθήκης επιλέγεται για την τρέχουσα επανάληψη. Σαν μέτρο αξιολόγησης για το συγκεκριμένο αλγόριθμο πρέπει να επιλεχθεί κάποιο που να ενσωματώνει την αγνότητα της συνθήκης και το βαθμό κάλυψης των παραδειγμάτων. Ένα τέτοιο μέτρο είναι το μέτρο Laplace αλλά και η κανονικοποιημένη εντροπία όπως παρουσιάστηκε στην παράγραφο 2.2.4.

Το δεύτερο έργο, όμοια αξιολογεί όλους τους πιθανούς συνδυασμούς των συνθηκών αυτών που έχουν ταυτιστεί στα παραδείγματα. Υπολογίζει τις κατανομές των κλάσεων και στη συνέχεια τις

αξιολογεί. Στο τέλος του δεύτερου έργου, ο οδηγός απλώς επιλέγει τον καλύτερο κανόνα από τη λίστα που προκύπτει από το δεύτερο έργο. Η επανάληψη σταματά με τον οδηγό να ελέγχει αν όλα τα παραδείγματα που καλύπτονται από τον τρέχων κανόνα υπερβαίνουν το όριο που έχει προσδιορίσει ο χρήστης και που παρέχεται ως παράμετρος.

Σε αυτό το βήμα, οι αλγόριθμοι κάλυψης συνεχίζουν αφαιρώντας τα παραδείγματα που καλύπτονται από τον επιλεγμένο κανόνα από το σύνολο των παραδειγμάτων. Σε ένα περιβάλλον με δεδομένα μεγάλης κλίμακας ωστόσο, αυτή η διαδικασία συχνά δεν είναι εφικτή, εφόσον τα παραδείγματα μπορεί να μην είναι υποκείμενα διαγραφής, εξαιτίας τις δυσκολίας συλλογής και εισαγωγής τις στο κατανεμημένο σύστημα αρχείων. Αν δεν υπάρχει τέτοιο θέμα, τότε η συστοιχία πρέπει να διαθέτει την κατάλληλη χωρητικότητα για να αποθηκεύσει τα ενδιάμεσα δεδομένα (τα ακάλυπτα παραδείγματα) που έχουν περίπου το διπλάσιο μέγεθος από το αυτό του συνόλου των παραδειγμάτων συν οποιαδήποτε αντιγραφή μπορεί να πραγματοποιεί η υλοποίηση τις Απεικόνισης/Μείωσης. Ο προτεινόμενος αλγόριθμος ακολουθεί μια διαφορετική προσέγγιση. Αντί να διαγράφει τα παραδείγματα που καλύπτονται σε κάθε επανάληψη, κατανέμει τη λίστα των κανόνων που έχουν βρεθεί τις εργασίες Μείωσης έτσι ώστε αυτές να αποφασίσουν αν το παράδειγμα που εξετάζουν έχει ήδη καλυφθεί ή όχι. Αυτή η διαδικασία είναι γνωστή στη βιβλιογραφία ως φιλτράρισμα.

Μετά την ολοκλήρωση τις επαναληπτικής διαδικασίας, αν παραμείνουν ακάλυπτα παραδείγματα ο οδηγός ξεκινά ένα μόνο έργο Απεικόνισης/Μείωσης για την εύρεση του τελικού κανόνα. Το έργο αυτό απλώς υπολογίζει τις κατανομές κλάσεων των υπόλοιπων παραδειγμάτων και επιλέγει την πλειοψηφούσα κατηγορία.

Το πρώτο έργο Απεικόνισης/Μείωσης κατασκευάζει μια λίστα από όλες τις συνθήκες που βρέθηκαν στο σύνολο παραδειγμάτων, συναθροίζει τις κατανομές κλάσεων και αξιολογεί κάθε συνθήκη. Και εδώ υπάρχει η απαίτηση να επιτευχθεί εκτενής και ταυτόχρονα ορθή κάλυψη των παραδειγμάτων.

Η εργασία της Απεικόνισης αρχικά λαμβάνει την τρέχουσα λίστα κανόνων από τον οδηγό για να αποφασίσει αν τα παραδείγματα που θα διαβάσει έχουν ήδη καλυφθεί. Εφόσον αυτή η λίστα είναι συνήθως μικρή, μπορεί να μεταφερθεί από την κατανεμημένη δευτερεύουσα μνήμη ή κάποιον παρόμοιο μηχανισμό διαμοιρασμού δεδομένων σε όλες τις εργασίες Απεικόνισης. Στην εργασία Απεικόνισης του έργου, η συνάρτηση απεικόνισης λαμβάνει ένα παράδειγμα τη φορά στη μορφή <κλειδί, τιμή> όπου το κλειδί αντιστοιχεί σε έναν προσδιοριστή και η τιμή αντιστοιχεί στο πραγματικό παράδειγμα. Ανάλογα με την τρέχουσα λίστα κανόνων, η συνάρτηση απεικόνισης αποφασίζει αν το παράδειγμα καλύπτεται και αν δεν καλύπτεται, συγκρατεί τις συνθήκες που βρέθηκαν στο τρέχον παράδειγμα μαζί με την κατηγορία του και τις προσθέτει σε ένα πίνακα.

Επίσης αυξάνει έναν καθολικό μετρητή για να προσδιορίσει το μέγεθος του συνόλου των παραδειγμάτων στην περίπτωση που δεν είναι γνωστός εκ των προτέρων.

Αλγόριθμος 2.5 - Η εργασία Απεικόνισης του πρώτου έργου χωρίς συνδυαστή

```
Είσοδος:      τρέχουσα λίστα κανόνων,
              παράδειγματα
Έξοδος:       <συνθήκη, κατανομή κλάσεων>

1.Συνάρτηση: Απεικόνιση (ID Παραδείγματος, Παράδειγμα, Συνθήκη, Κατανομή Κλ)
2.   Πλήθος στιγμιότυπων = πλήθος στιγμιότυπων + 1
3.   Αν το παράδειγμα δεν καλύπτεται από την τρέχουσα λίστα κανόνων τότε
4.      {συνθήκες}παραδείγματος = παραγωγή_συνθηκών(παράδειγμα)
5.      Για όλες τις συνθήκες στη λίστα {συνθήκες}παραδείγματος
6.         Κατανομή κλάσεων συνθήκης = {}
7.         Κατανομή κλάσεων = ενημέρωσε_συχνότητες (παράδειγμα)
8.     Εκπομπή(<συνθήκη, κατανομή_κλάσεων>)
9.  Τέλος επανάληψης
10.Τέλος επανάληψης
```

Το κλειδί εξόδου της εργασίας απεικόνισης, είναι σύνθετου τύπου και αντί να περιλαμβάνει τα δεδομένα καθαυτά περιλαμβάνει τους δείκτες της ιδιότητας και της τιμής. Η τιμή εξόδου όμοια είναι ένας σύνθετος τύπος δεδομένων αποτελούμενος από έναν πίνακα με τις κατανομές κλάσεων της αντίστοιχης συνθήκης του κλειδιού. Σε μια τυπική προσέγγιση Απεικόνισης/Μείωσης, σε αυτή τη φάση το ζευγάρι κλειδιού-τιμής θα είχε φυσιολογικά αποσταλεί στη συνάρτηση μείωσης για κάθε συνθήκη που βρέθηκε στο τρέχον παράδειγμα. Στην προτεινόμενη προσέγγιση αντίθετα, οι συνθήκες προστίθενται σε έναν συσχετιζόμενο πίνακα που συγκρατεί τις κατανομές κατηγοριών τους. Το ζευγάρι <συνθήκη, κατανομή κατηγοριών> κατά συνέπεια στέλνεται στη συνάρτηση μείωσης μόνο μετά την επεξεργασία όλων των τοπικών παραδειγμάτων.

Αυτή η τεχνική είναι γνωστή στη βιβλιογραφία ως "ενδοαπεικονικός συνδυασμός" (in mapper combiner) [56], αφού ενσωματώνει τη λειτουργικότητα ενός συνδυαστή μέσα στην Απεικόνιση και ουσιαστικά αποτελεί μια βελτιστοποίηση για τη μείωση των μηνυμάτων που ανταλλάσσονται ανάμεσα στην Απεικόνιση και τη Μείωση. Ο ψευδοκώδικας για την Απεικόνιση του πρώτου έργου χωρίς τη χρήση συνδυαστή φαίνεται στον Αλγόριθμος 2.5, ενώ με χρήση συνδυαστή φαίνεται στον Αλγόριθμος 2.6. Η επίδοση κάθε προσέγγισης αναλύεται στην ενότητα αξιολόγησης της συνολικής προσέγγισης.

Αλγόριθμος 2.6 - Η εργασία Απεικόνισης του πρώτου έργου με συνδυαστή

```
Είσοδος:      τρέχουσα λίστα κανόνων, παραδείγματα
Έξοδος:       <συνθήκη, κατανομή κατηγοριών>

1.Κατανομή κατηγοριών συνθηκών ← {}
2.Συνάρτηση Απεικόνιση(ID Παραδείγματος, Παράδειγμα, Συνθήκη, ΚατανομήΚατηγ)
3.  Πλήθος στιγμιότυπων = πλήθος στιγμιότυπων + 1
4.  Αν το παράδειγμα δεν καλύπτεται από την τρέχουσα λίστα κανόνων τότε
5.     {συνθήκες}παραδείγματος = παραγωγή_συνθηκών(παράδειγμα)
6.  Για όλες τις συνθήκες της λίστας {συνθήκες}παραδείγματος
7.     Αν η συνθήκη δεν έχει προστεθεί στις κατανομές κατηγοριών συνθηκών
8.    Πρόσθεσε τη συνθήκη στις κατανομές κατηγοριών συνθηκών
9.    Αύξησε τη συχνότητα της αντίστοιχης κατηγορίας της συνθήκης
10. Αλλιώς
11.    Αύξησε τη συχνότητα της αντίστοιχης κατηγορίας της συνθήκης
12.Τέλος συνάρτησης
13.
14.Συνάρτηση Ολοκλήρωση()
15.  Για κάθε συνθήκη στην κατανομή κατηγοριών συνθηκών
16.     εκπομπή(συνθήκη, κατανομή κατηγοριών)
17.Τέλος συνάρτησης
```

Στην εργασία Μείωσης, του έργου η συνάρτηση μείωσης λαμβάνει όλα τα ζευγάρια <Συνθήκη, Λίστα<κατανομή κατηγοριών>> από τις Απεικονίσεις, ένα κάθε φορά και τα συναθροίζει για να παράξει τις συνολικές κατανομές κατηγοριών για ολόκληρο το σύνολο παραδειγμάτων. Στο τέλος της συνάθροισης του κάθε παραδείγματος, η Μείωση αξιολογεί τη συγκεκριμένη συνθήκη, την ενσωματώνει στη συνθήκη και τη στέλνει στο κατανεμημένο σύστημα αρχείων. Ο Οδηγός τότε λαμβάνει την έξοδο και κατασκευάζει την τελική λίστα συνθηκών, την οποία θα ταξινομήσει, έτσι ώστε να επιλέξει τις k-καλύτερες συνθήκες. Ο ψευδοκώδικας για τη Μείωση του πρώτου έργου φαίνεται στον Αλγόριθμος 2.7.

Αλγόριθμος 2.7 - Η Μείωση του πρώτου έργου

```
Είσοδος <Συνθήκη, Λίστα<Κατανομή Κατηγοριών>>
Έξοδος <Συνθήκη>
1.Συνάρτηση Μείωση (<Συνθήκη, Λίστα<κατανομή κατηγοριών>, συνθήκη, null>)
2.  Για κάθε κατανομή κατηγοριών στη Λίστα<Κατανομή Κατηγοριών>
3.    Συνάθροισε τις κατανομές των αντίστοιχων κατηγοριών
4.  Αξιολόγησε τη συνθήκη
5.  Εκπομπή(συνθήκη)
```

Το δεύτερο έργο του αλγόριθμου είναι αρκετά όμοιο με το πρώτο. Στην προκειμένη περίπτωση, ο στόχος είναι να αξιολογήσει τους πιθανούς κανόνες που μπορούν να κατασκευαστούν από το σύνολο των καλύτερων συνθηκών που επιλέχθηκαν στο προηγούμενο έργο. Η Απεικόνιση κατά συνέπεια λαμβάνει α) την τρέχουσα λίστα κανόνων για να προσδιορίσει αν τα παραδείγματα είναι καλυμμένα β) τις k-καλύτερες συνθήκες από το προηγούμενο βήμα και γ) το μέγιστο πλήθος συνθηκών ανά κανόνα που έχει δοθεί από τον χρήστη σαν είσοδο στον αλγόριθμο.

Αλγόριθμος 2.8 - Η Απεικόνιση του δεύτερου έργου χωρίς συνδυαστή

```
Είσοδος:      τρέχουσα λίστα κανόνων,
              k-καλύτερες συνθήκες,
              μέγιστο πλήθος συνθηκών ανά κανόνα
Έξοδος:       <Κανόνας, Κατανομή Κλάσεων>
1.Συνάρτηση Απεικόνιση (<ID παραδείγματος, παράδειγμα,
2.                       Κανόνας, Κατανομή Κατηγοριών>)
3.  Αν το παράδειγμα δεν καλύπτεται από την τρέχουσα λίστα κανόνων τότε
4.    Ακάλυπτα παραδείγματα ← Ακάλυπτα παραδείγματα + 1
5.    Ταυτισμένες συνθήκες ← ταύτιση_συνθηκών(k-καλύτερες συνθήκες,
   παράδειγμα)
6.    Κανόνες με τις ταυτισμένες συνθήκες ← παραγωγή_συνδυασμών(συνθήκες)
7.       Για όλους τους κανόνες στους Κανόνες με τις ταυτισμένες συνθήκες
8.          Κατανομή κατηγοριών κανόνων ← {}
9.          Κατανομή κατηγοριών κανόνων ← ενημέρωσε_συνθήκες(παράδειγμα)
10.         Εκπομπή(κανόνας, κατανομή κατηγοριών)
```

Στο κομμάτι Απεικόνισης του έργου (ο ψευδοκώδικας του οποίου φαίνεται στον Αλγόριθμος 2.8 και τον Αλγόριθμος 2.9), ιδιαίτερο ενδιαφέρον παρουσιάζει η διαδικασία σύνθεσης κανόνων. Η συνάρτηση απεικόνιση αφού αποφασίσει αν το παράδειγμα έχει καλυφθεί ως τώρα ή όχι, προσδιορίζει ποιες από τις k-καλύτερες συνθήκες έχουν βρεθεί στο τρέχον παράδειγμα και κατασκευάζει κανόνες με όλους τους πιθανούς συνδυασμούς με μέγιστο μέγεθος αυτό που έχει προσδιορίσει ο χρήστης. Για παράδειγμα, έστω ότι σε ένα παράδειγμα με 10 ιδιότητες έχουν ταυτιστεί 3 μόνο συνθήκες από τις συνολικά 10 καλύτερες και παράλληλα ο χρήστης έχει ορίσει την κατασκευή κανόνων με μέγιστο μέγεθος 6 συνθήκες. Τότε η Μείωση θα συνθέσει όλους τους κανόνες τριών συνθηκών. Στην περίπτωση που στο παράδειγμα έχουν ταυτιστεί και οι 10 συνθήκες, η Μείωση θα συνθέσει όλους τους κανόνες με 6 συνθήκες. Η Απεικόνιση επίσης χρησιμοποιεί έναν "ενδοαπεικονικό συνδυαστή" για να στείλει μόνο τους κανόνες που έχουν ταυτιστεί στα τοπικά παραδείγματα.

Στο κομμάτι Μείωσης του δεύτερου έργου (ο ψευδοκώδικας φαίνεται στον Αλγόριθμος 2.10) η διαδικασία είναι ίδια με αυτήν του πρώτου έργου, εκτός από το γεγονός ότι στο τέλος ανατίθεται η πλειοψηφούσα κατηγορία που βρέθηκε στην κατανομή των κατηγοριών, στη συνέπεια του κανόνα.

Αλγόριθμος 2.9 - Η Απεικόνιση της δεύτερης εργασία με τη χρήση συνδυαστή

```
Είσοδος:      τρέχουσα λίστα κανόνων,
               k-καλύτερες συνθήκες,
               μέγιστο πλήθος συνθηκών ανά κανόνα
Έξοδος:       <Κανόνας, Κατανομή Κατηγοριών>
1. Κατανομές κατηγοριών κανόνων <κανόνας, κατανομή κατηγοριών> ← {}
2. Συνάρτηση Απεικόνιση (<ID παραδείγματος, παράδειγμα, Κανόνας, Κατανομή
   Κατηγοριών>)
3. Αν το παράδειγμα δεν καλύπτεται από την τρέχουσα λίστα κανόνων τότε
4.    Ακάλυπτα παραδείγματα ← Ακάλυπτα παραδείγματα + 1
5.    Ταυτισμένες συνθήκες ← ταύτιση_συνθηκών(k-καλύτερες συνθήκες,
   παράδειγμα)
6.    Κανόνες με τις ταυτισμένες συνθήκες ← παραγωγή_συνδυασμών(συνθήκες)
7.    Για όλους τους κανόνες στους Κανόνες με τις ταυτισμένες συνθήκες
8.       Αν ο κανόνας δεν υπάρχει στις κατανομές κατηγοριών κανόνων
9.           Κατανομές κανόνων ← κανόνας
10.          Αύξησε την αντίστοιχη συχνότητα της κατηγορίας του κανόνα
11.      Αλλιώς
12.          Αύξησε την αντίστοιχη συχνότητα της κατηγορίας του κανόνα
13.
14. Συνάρτηση Ολοκλήρωση()
15.  Για κάθε κανόνα στην κατανομή κατηγοριών κανόνων
16.     εκπομπή(κανόνας, κατανομή κατηγοριών)
17. Τέλος συνάρτησης
```

Αλγόριθμος 2.10 - Μείωση δεύτερου έργου

```
Είσοδος: <Κανόνας, Λίστα<Κατανομές Κατηγοριών>>
Έξοδος: <Κανόνας>
1.Συνάρτηση Μείωση(<Κανόνας, Λίστα<Κατανομές Κατηγοριών>, Κανόνας, Null>)
2.  Για όλες τις κατανομές κατηγοριών στη λίστα<Κατανομές κατηγοριών>
3.     Συνάθροισε τις αντίστοιχες συχνότητες των κατηγοριών
4.  Αξιολόγησε τον κανόνα
5.  Εκπομπή(Κανόνα)
```

Ο στόχος του τρίτου και τελευταίου έργου, είναι να βρεθεί ο προεπιλεγμένος κανόνας που καλύπτει τα υπόλοιπα ακάλυπτα παραδείγματα. Κάτι τέτοιο επιτυγχάνεται πολύ απλά υπολογίζοντας την κατανομή των κατηγοριών για όλα τα εναπομείναντα παραδείγματα και επιλέγοντας την κατηγορία με τη μέγιστη συχνότητα. Στη φάση Απεικόνισης η κατανομή των κατηγοριών υπολογίζεται παρόμοια με τα δυο προηγούμενα έργα. Όταν καλείται η συνάρτηση απεικόνισης σε ένα παράδειγμα, αυξάνεται η αντίστοιχη εγγραφή στον πίνακα που συγκρατεί τις κατανομές των

κατηγοριών. Στο τέλος της φάσης απεικόνισης, στέλνει τη μοναδική κατανομή κλάσεων. Στη φάση Μείωσης, οι κατανομές από όλες τις εργασίες Απεικόνισης συναθροίζονται για να παράξουν μια μοναδική καθολική κατανομή κλάσεων, από την οποία θα επιλεχθεί η πλειοψηφούσα κατηγορία και η οποία θα ανατεθεί στον κανόνα. Η τελική λίστα κανόνων ολοκληρώνεται με την προσθήκη αυτού του κανόνα.

2.5 Αξιολόγηση Επίδοσης Προτεινόμενης Προσέγγισης

Στην παράγραφο αυτή περιγράφεται η πειραματική αξιολόγηση των διαδικασιών της διακριτοποίησης και της επαγωγής κανόνων κατηγοριοποίησης. Η αξιολόγηση των διαδικασιών αυτών γίνεται από δυο οπτικές: α) την ακρίβεια γνωστών αλγόριθμων κατηγοριοποίησης αφού έχουν διακριτοποιηθεί με την προτεινόμενη προσέγγιση αλλά και την ακρίβεια του προτεινόμενου αλγόριθμου επαγωγής κανόνων σε επίλεκτα σύνολα δεδομένων και β) το κόστος επικοινωνίας μεταξύ των εργασιών Απεικόνισης και Μείωσης. Για τον αλγόριθμο επαγωγής κανόνων γίνεται μια επιπρόσθετη μελέτη για την παράλληλη επίδοσή του.

Για την αξιολόγηση των αλγόριθμων επαγωγής κανόνων ως προς την ακρίβειά τους στην κατηγοριοποίηση νέων στιγμιότυπων, συνήθως υπολογίζεται η αναλογία λαθών σε ένα νέο σύνολο στιγμιότυπων το οποίο δεν χρησιμοποιήθηκε στη διαδικασία κατασκευής τους. Αυτό το σύνολο είναι γνωστό ως σύνολο δοκιμών. Ωστόσο σε πολλές περιπτώσεις, ένα τέτοιο σύνολο δεν είναι διαθέσιμο ή είναι δύσκολο να παραχθεί. Σε τέτοιες περιπτώσεις, είτε διαχωρίζεται ένα κομμάτι των δεδομένων εκπαίδευσης και χρησιμοποιείται σαν σύνολο δοκιμών και το υπόλοιπο χρησιμοποιείται για την εκπαίδευση, είτε πολλαπλά κομμάτια του συνόλου εκπαίδευσης χρησιμοποιούνται - ένα κάθε φορά - για τη δοκιμή. Η πρώτη περίπτωση είναι γνωστή ως διαδικασία παρακράτησης (hold out), ενώ η δεύτερη είναι γνωστή ως σταυρωτή επικύρωση (cross validation). Στη δεύτερη περίπτωση, η ακρίβεια υπολογίζεται σαν ένας μέσος όρος της ακρίβειας που υπολογίστηκε σε κάθε πείραμα παρακράτησης.

Στο μοντέλο Απεικόνισης/Μείωσης, πέραν του συνηθισμένου κόστους στην επικοινωνία μέσα στη συστοιχία των κόμβων, υπάρχει επίσης κόστος ανάμεσα στις Απεικονίσεις και τις Μειώσεις που αφορά τον τρόπο υλοποίησής τους. Τρία επιπρόσθετα κόστη λαμβάνονται υπόψη στα έργα Απεικόνισης/Μείωσης α) το κόστος εκτέλεσης της εργασίας που τρέχει στην Απεικόνιση β) το κόστος επικοινωνίας για τη μεταφορά δεδομένων από τις Απεικονίσεις στις Μειώσεις και γ) το κόστος εκτέλεσης της εργασίας που τρέχει στη Μείωση. Επειδή οι Μειώσεις τρέχουν στους κόμβους όπου υπάρχουν τα δεδομένα που πρέπει να επεξεργαστούν, δεν υπάρχει αρχικό επικοινωνιακό κόστος, μόνο το κόστος εκτέλεσης της συνάρτησης απεικόνισης. Η Απεικόνιση ωστόσο στέλνει ζευγάρια κλειδιού-τιμής που πρέπει να μεταφερθούν στους κόμβους που τρέχουν τις εργασίες Μείωσης. Το κόστος της μεταφοράς ζευγαριών κλειδιού-τιμής από τις Απεικονίσεις στις Μειώσεις,

είναι ανάλογο του μεγέθους των ζευγαριών που στέλνουν οι Απεικονίσεις. Αυτό το κόστος στην επικοινωνία εκφράζεται με τον όρο ρυθμός αντιγραφής (replication rate) που είναι η μέση επικοινωνία μεταξύ των Απεικονίσεων και των Μειώσεων και πιο συγκεκριμένα είναι ο μέσος όρος των ζευγαριών κλειδιού τιμής που δημιουργούν οι Απεικονίσεις για το σύνολο των δεδομένων εισόδου. Τέλος το κόστος εκτέλεσης της εργασίας μείωσης, ουσιαστικά ορίζεται από το μέγεθος της λίστας των τιμών που λαμβάνει κάθε συνάρτηση μείωσης για ένα κλειδί. Αυτό το μέγεθος είναι γνωστό ως μέγεθος μείωσης (reducer size) [58].

Για την αξιολόγηση των παράλληλων ή κατανεμημένων αλγόριθμων μηχανικής μάθησης ωστόσο, η ακρίβεια δεν είναι αρκετή. Δυο επιπρόσθετα μέτρα χρησιμοποιούνται, που δείχνουν τη δυνατότητα του αλγόριθμου να μειώνει το χρόνο εκτέλεσης όταν προστίθενται περισσότεροι υπολογιστικοί κόμβοι, ή η δυνατότητα χειρισμού μεγαλύτερων συνόλων δεδομένων με την πρόσθεση περισσότερων μηχανών. Αυτά τα μέτρα είναι γνωστά ως "speedup" και "scaleup" αντίστοιχα. Ένα τρίτο μέτρο που χρησιμοποιείται συχνά είναι το "sizeup" που μετρά πόσο μεγαλύτερη είναι η εκτέλεση σε ένα δεδομένο σύστημα, όταν το μέγεθος του συνόλου δεδομένων αυξάνεται σταθερά. Ο ιδανικός παράλληλος αλγόριθμος επιδεικνύει γραμμικό speedup: ένα σύστημα με p περισσότερους κόμβους θα επιδείξει ένα speedup p. Ωστόσο, το γραμμικό speedup είναι δύσκολο να επιτευχθεί γιατί το κόστος στην επικοινωνία αυξάνεται όσο το πλήθος των κόμβων μεγαλώνει. Κατά συνέπεια ένα άλλο θέμα που πρέπει να εξεταστεί στους παράλληλους αλγόριθμους είναι το επικοινωνιακό κόστος.

2.5.1 Ακρίβεια κατηγοριοποίησης

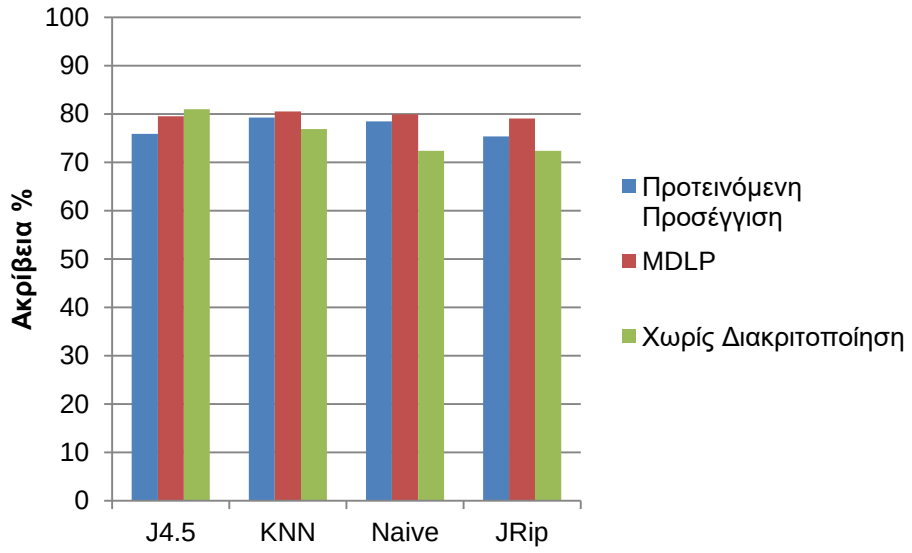
2.5.1.1 Ακρίβεια κατηγοριοποίησης με τη διαδικασία διακριτοποίησης

Η ακρίβεια της προτεινόμενης προσέγγισης διακριτοποίησης μετρήθηκε χρησιμοποιώντας οχτώ δημοσίως διαθέσιμα συνεχή σύνολα δεδομένων τα οποία επεξεργάστηκαν τέσσερις γνωστοί αλγόριθμοι κατηγοριοποίησης αφού προηγουμένως τα σύνολα αυτά είχαν διακριτοποιηθεί α) με την προτεινόμενη προσέγγιση και β) με την προσέγγιση MDLP [59] που θεωρείται από τους καλύτερους αλγόριθμους διακριτοποίησης [43]. Οι αλγόριθμοι επίσης δοκιμάστηκαν με τις προεπιλεγμένες ρυθμίσεις τους χωρίς να έχει γίνει κάποια διακριτοποίηση προηγουμένως. Τα χαρακτηριστικά των συνόλων δεδομένων που χρησιμοποιήθηκαν για την αξιολόγηση αυτή φαίνονται στον Πίνακας 2.1, ενώ τα σύνολα αυτά είναι διαθέσιμα στο αποθετήριο UCI [60]. Οι αλγόριθμοι που επιλέχθηκαν για τη μελέτη ήταν οι J4.5, k-nearest neighbor, Naïve Bayes και JRip, χρησιμοποιώντας τις υλοποιήσεις του πακέτου λογισμικού WEKA [61].

Πίνακας 2.1 - Τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την αξιολόγηση της ακρίβειας μετά από τη διακριτοποίηση

	Στιγμιότυπα	**Ιδιότητες**	**Κατηγορίες**
Australian [62]	690	14	2
Ecoli [63]	336	8	8
Glass [64]	214	10	7
Haberman [65]	306	3	2
Segment [66]	2310	19	7
Sonar [67]	208	60	2
Vehicle [68]	946	18	4
Waveform [69]	5000	21	3

Στην Εικόνα 2.3 φαίνεται ο μέσος όρος της ακρίβειας των τεσσάρων αλγόριθμων κατηγοριοποίησης που χρησιμοποιήθηκαν στα σύνολα δεδομένων του Πίνακας 2.1. Μια πρώτη παρατήρηση προκύπτει από τη σύγκριση της ακρίβειας των αλγορίθμων με διακριτοποίηση και χωρίς διακριτοποίηση. Για τους αλγόριθμους KNN, Naïve Bayes και JRip, η χρήση διακριτοποίησης βελτίωσε την ακρίβειά τους από δύο έως τέσσερις ποσοστιαίες μονάδες, ενώ αντίθετα η χρήση διακριτοποίησης μείωσε την ακρίβεια του J4.5 κατά μια έως τρεις μονάδες, κάτι που σημαίνει ότι η εσωτερική, δυναμική διακριτοποίησή του τελευταίου υπερτερεί έναντι των δυο προσεγγίσεων.

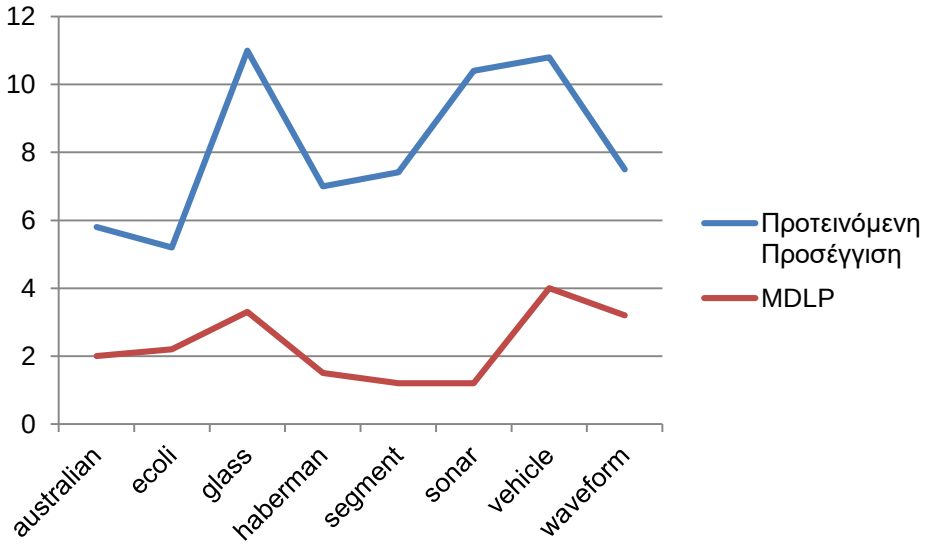


Εικόνα 2.3 - Σύγκριση διακριτοποιήσεων για διάφορους αλγόριθμους κατηγοριοποίησης

Συγκρίνοντας την προσέγγιση MDLP με την προτεινόμενη προσέγγιση, είναι εμφανής η υπεροχή της πρώτης σε όλα τα σύνολα δεδομένων και με όλους τους δοκιμασμένους αλγόριθμους, με μια διαφορά που κυμαίνεται από δύο έως τέσσερις ποσοστιαίες μονάδες. Η διαφορά αυτή ωστόσο συνοδεύεται από περισσότερο κόστος στη μνήμη και την ανάγκη εκτέλεσης πολλαπλών επαναλήψεων. Στον τομέα αυτό η προτεινόμενη προσέγγιση υπερτερεί, αφού δεν χρειάζεται να φορτώσει όλα τα δεδομένα στη μνήμη και εξάγει τα αποτελέσματα με ένα πέρασμα στα δεδομένα,

κάτι που σε συνδυασμό με το γεγονός ότι εκμεταλλεύεται πλήρως το μοντέλο της Απεικόνισης/Μείωσης, οδηγεί στο συμπέρασμα ότι μπορεί να αποκριθεί στην κλιμάκωση των δεδομένων.

Μια δεύτερη μέτρηση που μπορεί να εξαχθεί από τα αποτελέσματα της εκτέλεσης των τεσσάρων αλγόριθμων κατηγοριοποίησης στα σύνολα δεδομένων του Πίνακας 2.1, αφορά το πλήθος των διαστημάτων που κατασκευάζει κάθε μέθοδος διακριτοποίησης. Στην Εικόνα 2.4 φαίνεται ο μέσος όρος διαστημάτων που παρήχθησαν από τις δυο προσεγγίσεις διακριτοποίησης σε κάθε σύνολο δεδομένων.



Εικόνα 2.4 - Σύγκριση πλήθους διαστημάτων προτεινόμενης προσέγγισης με την MDLP

Από την εικόνα είναι ορατή η υπεροχή της προσέγγισης MDLP, αφού παράγει κατά μέσο όρο δυο έως τρία διαστήματα, σε αντίθεση με την προτεινόμενη προσέγγιση που παράγει έξι έως έντεκα. Η διαφορά αυτή οφείλεται στο γεγονός ότι ο προτεινόμενος αλγόριθμος ενεργεί σε τοπικά δεδομένα σε αντίθεση με τον MDLP που ενεργεί καθολικά σε όλα τα δεδομένα, αλλά και στην απλότητα του κριτηρίου ενσωμάτωσης δυο διαδοχικών διαστημάτων (δυο διαδοχικά διαστήματα ενσωματώνονται μόνο αν έχουν την ίδια κατηγορία). Και αυτή η υπεροχή ωστόσο συνοδεύεται από κόστος στη μνήμη και σε επαναλήψεις. Στο σημείο αυτό να τονιστεί ότι ένα μεγάλο πλήθος διαστημάτων δεν επηρεάζει απαραίτητα αρνητικά την ακρίβεια ενός αλγόριθμου κατηγοριοποίησης. Αντίθετα μπορεί να αποβεί επωφελής σε περιπτώσεις π.χ. που αναζητούνται ακραίες τιμές (outliers), όπως στην ανίχνευση ανωμαλιών (intrusion detection). Ένα μεγάλο πλήθος διαστημάτων ωστόσο σίγουρα επηρεάζει τον χρόνο εκτέλεσης ενός αλγόριθμου κατηγοριοποίησης, αφού θα πρέπει να εξεταστούν περισσότερα διαστήματα.

2.5.1.2 Ακρίβεια κατηγοριοποίησης για την προτεινόμενη προσέγγιση επαγωγής κανόνων

Η επίδοση της προτεινόμενης προσέγγισης μετρήθηκε χρησιμοποιώντας διάφορα δημοσίως διαθέσιμα σύνολα με ένα περιορισμένο πλήθος στιγμιότυπων και ιδιοτήτων απ' όπου μετρήθηκε η ακρίβειά της και συγκρίθηκε με αντίστοιχη ακρίβεια ενός συνόλου γνωστών αλγόριθμων επαγωγής κανόνων και κατασκευής δέντρων κατηγοριοποίησης. Τα σύνολα δεδομένων είναι διαθέσιμα στο αποθετήριο UCI [60] και οι υλοποιήσεις των αλγόριθμων αυτών παρέχονται από το πακέτο λογισμικού WEKA [61]. Τα χαρακτηριστικά των συνόλων που χρησιμοποιήθηκαν σε αυτόν τον γύρο των επαναλήψεων παρουσιάζονται στον Πίνακας 2.2.

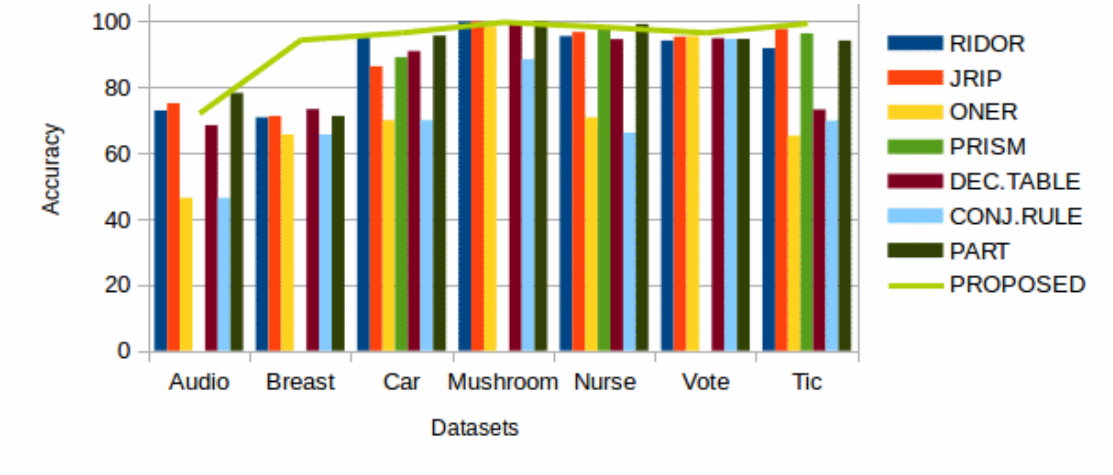
Πίνακας 2.2 - Τα χαρακτηριστικά των συνόλων δεδομένων που χρησιμοποιήθηκαν για την αξιολόγηση

	Στιγμιότυπα	**Ιδιότητες**	**Κατηγορίες**	**Συνθήκες**	**Ελλείπουσες τιμές**
Audiology	226	69	24	154	ναι
Breast Cancer	699	10	2	100	ναι
Car	17218	6	4	25	όχι
Mushroom	8124	22	2	126	όχι
Nursery	12960	7	5	24	όχι
Vote	435	16	2	48	ναι
Tic Tac Toe	958	9	2	27	όχι

Ο αλγόριθμος υλοποιήθηκε στη Java με το Hadoop 1.2.1 API και για τους στόχους του πειράματος αυτού δοκιμάστηκε σε ρύθμιση αυτονομίας (standalone mode). Για όλα τα πειράματα ο αλγόριθμος έτρεχε με την τιμή 0.01% για την παράμετρο "ποσοστό των ακάλυπτων παραδειγμάτων" που ουσιαστικά σημαίνει ότι καλύφθηκαν όλα παραδείγματα. Η παράμετρος "μέγιστο πλήθος συνθηκών ανά κανόνα" είχε τεθεί στη μέγιστα υποστηριζόμενη τιμή ανά σύνολο (εκτός από τα σύνολα Audiology και Mushroom όπου τέθηκε στο μισό του πλήθους των ιδιοτήτων). Οι υπόλοιποι από τους αλγόριθμους χρησιμοποιήθηκαν με τις προεπιλεγμένες ρυθμίσεις τους.

Τα αποτελέσματα που συγκρίνουν την ακρίβεια του αλγόριθμου με αυτήν των άλλων προσεγγίσεων με μια 10πλή σταυρωτή επικύρωση φαίνονται στην Εικόνα 2.5. Όπως φαίνεται στην εικόνα, η προτεινόμενη προσέγγιση ανακάλυψε κανόνες με ακρίβεια πολύ κοντά στην ακρίβεια των υπόλοιπων αλγόριθμων για όλα τα σύνολα δεδομένων.

Τα επόμενα αποτελέσματα της αξιολόγησης της προσέγγισης αφορούν την απλότητα της λίστας των κανόνων που ανακαλύφθηκαν, η οποία στη βιβλιογραφία μετριέται από πλήθος των κανόνων που βρέθηκαν και το μέσο πλήθος των συνθηκών ανά κανόνα. Τα αποτελέσματα φαίνονται στον Πίνακας 2.3. Πρέπει να σημειωθεί ότι για κάποιος αλγόριθμους δεν παρέχεται αυτή η πληροφορία από το WEKA (περιγράφεται ως N/A).



Εικόνα 2.5- Σύγκριση ακρίβειας κατηγοριοποίησης προτεινόμενης προσέγγισης με γνωστούς αλγόριθμους επαγωγής κανόνων

Πίνακας 2.3 - Πλήθος των συνθηκών ανά κανόνα

	Προτ.	JRip	OneR	Prism	Dec. Tabl	C. Rule	PART	RIDOR
Audio	51(1.1)	18(1.7)	3(1)	N/A	27(N/A)	N/A	21(2.4)	105(1.29)
Breast	45(1.9)	3(2)	13(1)	N/A	25(N/A)	N/A	20(2.1)	3(2.5)
Car	56(2.1)	49(3.9)	4(1)	247(5.4)	432(N/A)	N/A	68(2.4)	124(1.25)
Mushr	17(1)	9(1.1)	9(1)	N/A	167(N/A)	N/A	13(1.5)	12(1.6)
Nurse	97(2.6)	131(4.6)	3(1)	587(5.5)	810(N/A)	N/A	220(3.1)	192(2.3)
Vote	20(2.6)	4(2)	4(1)	N/A	40(N/A)	N/A	7(1.8)	3(2)
Tic	61(2.5)	9(3)	3(1)	56(3.6)	190(N/A)	N/A	49(2.7)	8(3)

2.5.2 Επικοινωνιακό κόστος

Σε αυτήν την παράγραφο γίνεται μια ανάλυση του επικοινωνιακού κόστους που έχουν τα έργα Απεικόνισης/Μείωσης. Ξεκινώντας με τη διαδικασία της διακριτοποίησης, αρχικά για όλα τα παραδείγματα του τμήματος που αναλύει μια Απεικόνιση, δημιουργείται ένα διάστημα το οποίο θα πρέπει να συγχωνευτεί με κάποιο από τα B υπάρχοντα διαστήματα. Το νέο διάστημα συγκρίνει το σταθμισμένο μέσο του (που ουσιαστικά είναι η ίδια η τιμή του, αφού περιέχει μόνο ένα παράδειγμα) με όλους τους σταθμισμένους μέσους όλων των διαστημάτων, και συγχωνεύεται με το πιο κοντινό διάστημα. Στο τέλος της συνάρτησης της Απεικόνισης, όταν θα έχουν εξεταστεί όλα τα παραδείγματα του τοπικού τμήματος του συνόλου δεδομένων, μεταδίδονται ακριβώς B στο πλήθος διαστήματα, επομένως ο ρυθμός αντιγραφής είναι σταθερός και ισούται με B. Όντας σταθερός ο ρυθμός αντιγραφής το μέγεθος μείωσης θα είναι B*M, όπου M το πλήθος των Απεικονίσεων που εκτελέστηκαν κατά τη διάρκεια του έργου.

Η προτεινόμενη προσέγγιση επαγωγής κανόνων αποτελείται από δυο έργα. Στην Απεικόνιση της εργασίας του πρώτου έργου γίνεται έλεγχος αν το τρέχον παράδειγμα έχει ήδη καλυφθεί. Αυτό το βήμα αποτελείται από το ταίριασμα καθενός από τους r κανόνες που βρίσκονται στην τρέχουσα λίστα (μέχρι ένας από αυτούς να ταιριάξει με το τρέχον παράδειγμα) καθένας από τους οποίους έχει

n συνθήκες. Κατά συνέπεια το ταίριασμα έχει πολυπλοκότητα O(rn). Η διαδικασία εύρεσης των συνθηκών του τρέχοντος παραδείγματος είναι η ίδια με αυτή της εξόδου της συνθήκης ή της εισαγωγής της στον πίνακα. Κατά συνέπεια αυτό το βήμα έχει πολυπλοκότητα O(n), όπου n είναι το πλήθος των συνθηκών που βρίσκονται στο παράδειγμα, το οποίο είναι το ίδιο με το πλήθος των ιδιοτήτων μείον ένα (την κατηγορία).

Σχετικά με την αναλογία αντιγραφής του πρώτου έργου, είναι δυνατός ο υπολογισμός της μέγιστης τιμής που μπορεί να ληφθεί για κάθε παράδειγμα. Αν δεν γίνει χρήση της βελτιστοποίησης του συνδυαστή, κάθε παράδειγμα θα έχει σαν έξοδο ακριβώς k ζευγάρια κλειδιού τιμής, όπου k είναι το πλήθος των συνθηκών που βρέθηκαν στο παράδειγμα. Το k ουσιαστικά ισούται με το πλήθος των ιδιοτήτων του συνόλου δεδομένων. Κατά συνέπεια, ο ρυθμός αντιγραφής είναι σταθερός και ισούται με k. Αν χρησιμοποιηθεί συνδυαστής ωστόσο, οι συνθήκες δεν θα εξαχθούν αλλά θα προστεθούν σε ένα πίνακα. Στο τέλος της εργασίας Απεικόνισης, το πλήθος των ζευγαριών κλειδιού τιμής που θα εξαχθούν θα είναι ακριβώς το ίδιο με τον αριθμό όλων των συνθηκών που βρέθηκαν στο τοπικό σύνολο δεδομένων που επεξεργάστηκε η Απεικόνιση (δηλαδή το μέγεθος του πίνακα). Στη χειρότερη περίπτωση θα είναι το πλήθος όλων των πιθανών συνθηκών που βρίσκονται στο σύνολο δεδομένων. Το μέγεθος σε bytes των ζευγαριών κλειδιού τιμής κρατιέται στο ελάχιστο δυνατό. Ουσιαστικά αποτελεί το σύνολο των δεικτών που δείχνουν στις πραγματικές τιμές των ιδιοτήτων που βρέθηκαν στο σύνολο δεδομένων. Το πλήθος του συνόλου των δεικτών εξαρτάται από το πλήθος των τιμών της κατηγορίας, αφού αυτό είναι που καθορίζει την τιμή της κατανομής της κατηγορίας του ζευγαριού κλειδιού-τιμής της εξόδου.

Το μέγεθος μείωσης όταν δεν χρησιμοποιείται συνδυαστής, αποτελείται από τη συχνότητα εμφάνισης της συνθήκης στο σύνολο δεδομένων, που στη χειρότερη περίπτωση είναι το μέγεθος του συνόλου δεδομένων, αφού μια συνθήκη δεν μπορεί να εμφανίζεται περισσότερες από μια φορές σε ένα παράδειγμα. Όταν χρησιμοποιείται ένας συνδυαστής ωστόσο, το μέγεθος μείωσης θα είναι όσο το πλήθος των τμημάτων που δίνονται στις Απεικονίσεις, αφού κάθε Απεικόνιση θα έχει σαν έξοδο κάθε συνθήκη ακριβώς μια φορά.

Η Απεικόνιση του δεύτερου έργου, επίσης αρχικά ελέγχει αν το παράδειγμα που εξετάζεται την τρέχουσα στιγμή έχει ήδη καλυφτεί. Η διαδικασία εύρεσης των κανόνων ωστόσο, είναι σχετικά πιο πολύπλοκη από αυτήν της εύρεσης των συνθηκών του πρώτου έργου. Αρχικά η συνάρτηση ταύτιση συνθηκών (γραμμή 5) ελέγχει αν οι k καλύτερες συνθήκες που έχουν επιλεχθεί από το πρώτο έργο έχουν βρεθεί στο τρέχον παράδειγμα. Η διαδικασία αυτή έχει πολυπλοκότητα O(kn), όπου n είναι το πλήθος των συνθηκών που έχουν βρεθεί στο παράδειγμα. Στη συνέχεια η συνάρτηση παραγωγή συνδυασμών (γραμμή 6) υπολογίζει όλους τους πιθανούς συνδυασμούς των συνθηκών που έχουν βρεθεί στο προηγούμενο βήμα. Αφού είναι αναγκαίοι κανόνες που αποτελούνται από μια συνθήκη σε πλήθος μέχρι την τιμή της μεταβλητής μέγιστο πλήθος συνθηκών ανά κανόνα, το πλήθος τους μπορεί να υπολογιστεί ως εξής:

$$\sum_{i=1}^{n} \frac{n!}{i!\,(n-1)!} = 2^n - 1 \tag{2.1}$$

Όπου n το μέγιστο πλήθος συνθηκών ανά κανόνα.

Κατά συνέπεια, αυτό το βήμα έχει O(2^n)πολυπλοκότητα, που ουσιαστικά αποτελεί την αναλογία αντιγραφής αν δεν χρησιμοποιηθεί συνδυαστής. Αν χρησιμοποιηθεί ωστόσο, η αναλογία αντιγραφής θα είναι ακριβώς ίδια με το πλήθος των κανόνων που έχουν βρεθεί στο τοπικό σύνολο δεδομένων που επεξεργάστηκε η Απεικόνιση.

Το μέγεθος μείωσης όταν δεν χρησιμοποιείται συνδυαστής, είναι ουσιαστικά η συχνότητα εμφάνισης ενός κανόνα στο σύνολο δεδομένων, που όμοια με τις συνθήκες, στη χειρότερη περίπτωση μπορεί να φτάσει το μέγεθος του συνόλου δεδομένων. Επίσης, και εδώ όταν χρησιμοποιείται συνδυαστής το μέγεθος μείωσης θα είναι όσο μεγάλο όσο ο αριθμός των τμημάτων εισόδου που έχουν αποδοθεί στις Απεικονίσεις, αφού κάθε Απεικόνιση θα έχει σαν έξοδο κάθε κανόνα ακριβώς μια φορά για όλο το τμήμα εισόδου.

2.5.3 Παράλληλη Επίδοση

Ήδη παρουσιάστηκε ο ορισμός των όρων "speedup", "scaleup" και "sizeup" και τώρα δίνεται ο ορισμός των αντίστοιχων τύπων τους.

Το speedup ενός κατανεμημένου συστήματος με n κόμβους ορίζεται ως:

$$Speedup(n) = \frac{T_1}{T_n} \tag{2.2}$$

όπου T_1είναι ο υπολογιστικός χρόνος που αντιστοιχεί σε ένα κόμβο και T_nείναι ο υπολογιστικός χρόνος για n κόμβους.

Το scaleup ενός κατανεμημένου συστήματος με n κόμβους υπολογίζεται ως

$$Scaleup(n) = \frac{T_d^1}{T_{nxd}^n} \tag{2.3}$$

όπου T_d^1 είναι ο υπολογιστικός χρόνος για την επεξεργασία d δεδομένων σε έναν κόμβο και T_{nxd}^n είναι ο χρόνος που χρειάζεται για τον υπολογισμό n φορές περισσότερων δεδομένων σε n κόμβους.

Τέλος το sizeup ενός κατανεμημένου υπολογιστικού συστήματος με n κόμβους ορίζεται ως:

$$Sizeup(n) = \frac{T_{nxd}}{T_d} \tag{2.4}$$

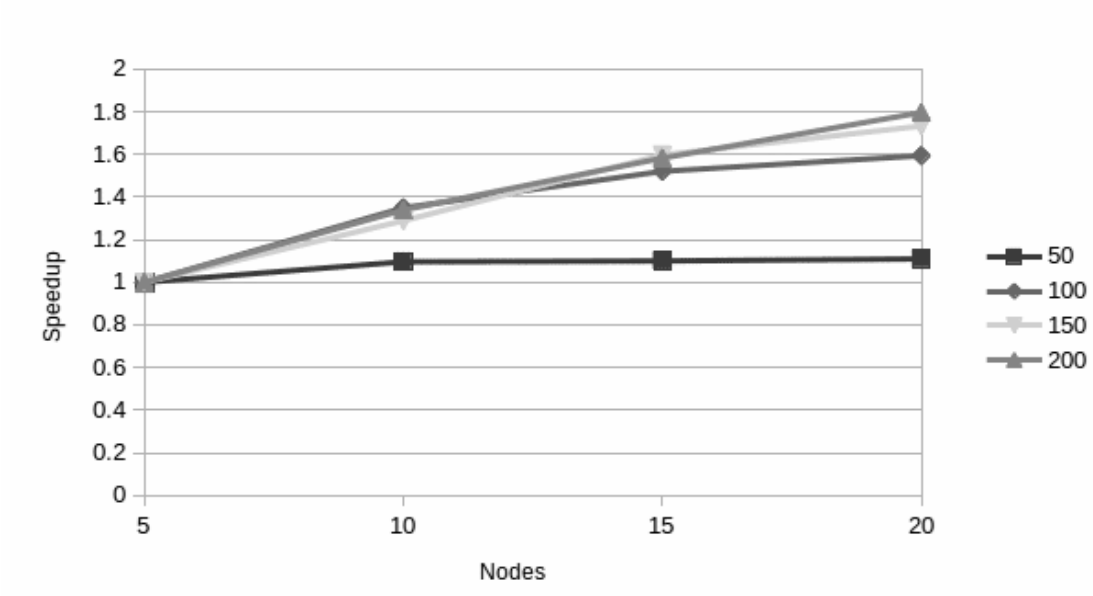
όπου T_{nxd} είναι ο υπολογιστικός χρόνος που χρειάζεται για την επεξεργασία ενός συνόλου δεδομένων n φορές μεγαλύτερο από το αρχικό και T_d είναι ο υπολογιστικός χρόνος για την επεξεργασία του αρχικού συνόλου δεδομένων.

Για να αξιολογηθεί η παράλληλη επίδοση της προτεινόμενης προσέγγισης επαγωγής κανόνων κατηγοριοποίησης, χρησιμοποιήθηκε μια συστοιχία μεταβλητού μεγέθους στην υπηρεσία Amazon Elastic MapReduce [70] που αποτελείται από 5 μέχρι 20 εικονικές μηχανές του τύπου M2.xlarge που παρέχουν 2 εικονικούς επεξεργαστές, 17 GB μνήμης και 420GB αποθηκευτικού χώρου ανά μηχανή. Για τις μετρήσεις αυτές χρησιμοποιήθηκαν συνθετικά σύνολα δεδομένων, τα χαρακτηριστικά των οποίων φαίνονται στον Πίνακας 2.4.

Πίνακας 2.4- Χαρακτηριστικά των συνθετικών συνόλων δεδομένων που χρησιμοποιήθηκαν σε πειράματα

Όνομα	**Παραδείγματα**	**Ιδιότητες**	**Κατηγορίες**	**Πιθανές συνθήκες**
50M10A5V	50.000.000	10	5	1000
100M10A5V	100.000.000	10	5	1000
150M10A5V	150.000.000	10	5	1000
200M10A5V	200.000.000	10	5	1000

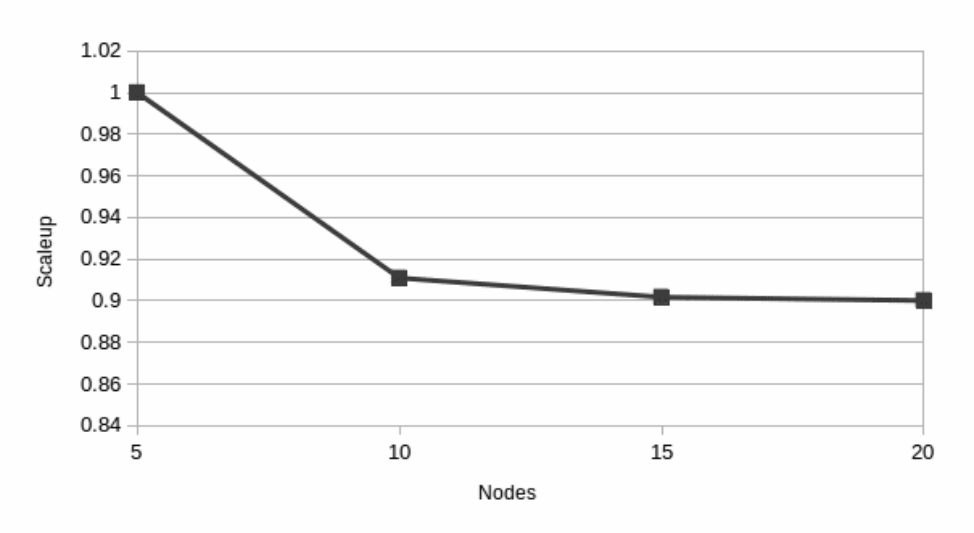
Όπως φαίνεται στην Εικόνα 2.6, με την αύξηση του μεγέθους του συνόλου δεδομένων, το speedup του αλγόριθμου τείνει να είναι περίπου γραμμικό, ιδιαίτερα για τα μεγαλύτερα δεδομένα. Όσο μεγαλύτερο είναι το μέγεθος του συνόλου δεδομένων, τόσο μεγαλύτερο πρέπει να είναι το speedup και αυτό παρατηρείται στα ευρήματα. Μια επιπρόσθετη παρατήρηση είναι ότι για μικρότερα σύνολα δεδομένων, δεν πετυχαίνεται υψηλό speedup όσο αυξάνουν οι κόμβοι της συστοιχίας. Αυτό δείχνει ότι οι επιπρόσθετοι κόμβοι δεν χρησιμοποιούνται κατά την εκτέλεση και κατά συνέπεια το ιδανικό μέγεθος της ιδανικής συστοιχία για αυτό το σύνολο δεδομένων είναι 10.



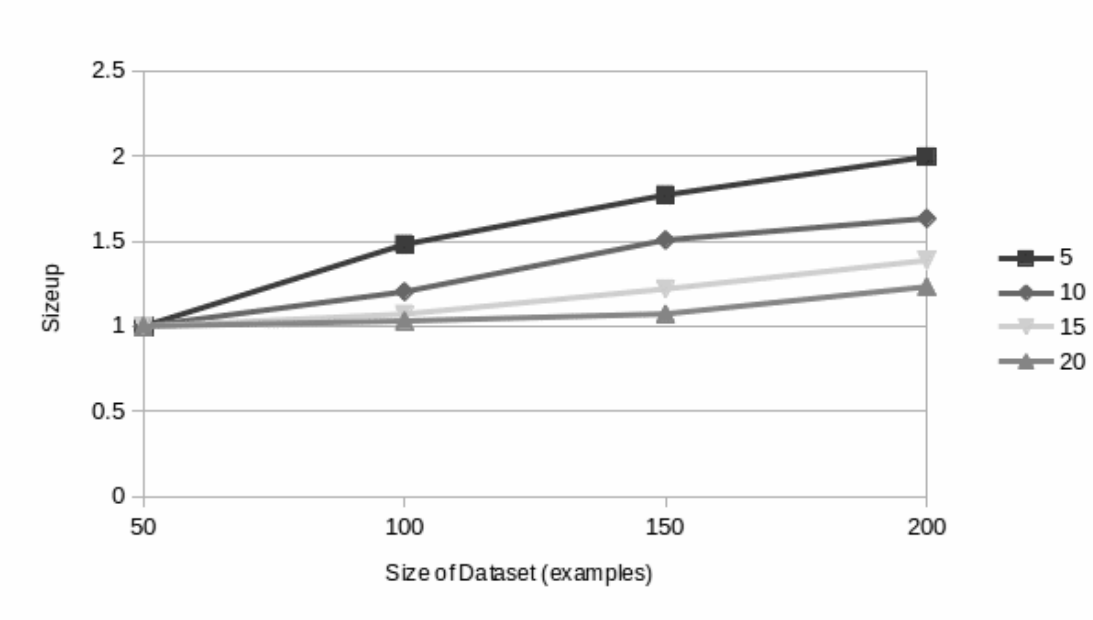
Εικόνα 2.6 - Speedup προτεινόμενης προσέγγισης

Για την αξιολόγηση της προτεινόμενης προσέγγισης ως προς το scaleup, αυξήθηκε το μέγεθος του συνόλου δεδομένων αναλογικά με την αύξηση των κόμβων στη συστοιχία. Συνεπώς, τα σύνολα δεδομένων με 50, 100, 150 και 200 παραδείγματα χρησιμοποιήθηκαν σε μια συστοιχία από 5, 10, 15 και 20 μηχανές. Η επίδοση ως προς το scaleup του αλγόριθμου φαίνεται στην Εικόνα 2.7. Θεωρητικά, η ιδανική συστοιχία υπολογιστών θα είχε ένα σταθερό scaleup ίσο με 1, ωστόσο στην πράξη κάτι τέτοιο δεν είναι δυνατό. Γενικά το scaleup ενός αλγόριθμου παρουσιάζει μια σταδιακή πτώση με την αύξηση του μεγέθους του συνόλου δεδομένων και το μέγεθος της συστοιχίας. Όπως φαίνεται στην Εικόνα 2.7, το scaleup για τον προτεινόμενο αλγόριθμο μειώνεται γρήγορα και όταν το σύνολο δεδομένων μεγαλώνει, μειώνεται πιο αργά. Σε συνδυασμό με το γεγονός ότι το scaleup κρατείται πάνω από την τιμή 0.9 μπορούμε να πούμε ότι ο αλγόριθμος πετυχαίνει καλή κλιμάκωση.

Τέλος, για να αξιολογηθεί ο δείκτης sizeup του αλγόριθμου, πραγματοποιήθηκε μια σειρά πειραμάτων σε 4 συστοιχίες κρατώντας τον αριθμό των κόμβων σταθερό και χρησιμοποιώντας τα σύνολα δεδομένων από 50, 100, 150 και 200 εκατομμύρια παραδείγματα. Τα αποτελέσματα μπορούν να φανούν στην Εικόνα 2.8. Συνήθως ένα γραμμικό sizeup θεωρείται η βάση και κάθε τιμή κάτω από αυτή θεωρείται καλή. Τα αποτελέσματα που λήφθηκαν, δείχνουν ότι ακόμα και για ένα σύνολο τέσσερις φορές μεγαλύτερο από το αρχικό, το sizeup ίσα ίσα που έχει διπλασιαστεί. Κάτι τέτοιο δείχνει ότι ο αλγόριθμος πετυχαίνει αρκετά καλό sizeup.



Εικόνα 2.7 - Scaleup προτεινόμενης προσέγγισης



Εικόνα 2.8 - Sizeup προτεινόμενης προσέγγισης

2.6 Συμπεράσματα

Στην ενότητα αυτή παρουσιάστηκε ένας αλγόριθμος διακριτοποίησης και ένας αλγόριθμος κάλυψης για την κατασκευή κανόνων κατηγοριοποίησης σε μεγάλα σύνολα δεδομένων με την τεχνική της Απεικόνισης/Μείωσης. Και οι δυο προσεγγίσεις σχεδιάστηκαν λαμβάνοντας υπόψη τα ιδιαίτερα χαρακτηριστικά του μοντέλου προγραμματισμού Απεικόνισης/Μείωσης. Απεδείχθη ότι με τη χρήση δυο απλών έργων Απεικόνισης/Μείωσης είναι δυνατή η αναζήτηση στο χώρο των ιδιοτήτων-τιμών και η κατασκευή κανόνων αναζητώντας στους συνδυασμούς τους, με ένα τρόπο κλιμακωτό ως προς το μέγεθος του συνόλου παραδειγμάτων. Ο αλγόριθμος κατηγοριοποίησης ωστόσο έχει σαν προαπαιτούμενο ότι τα δεδομένα είναι διακριτά.

Κατά συνέπεια, σε περιπτώσεις με αριθμητικά δεδομένα, είναι απαραίτητο να προηγηθεί μια διαδικασία διακριτοποίησης. Η προτεινόμενη προσέγγιση της παραγράφου 2.4.1, αποτελεί μια απλή και κλιμακούμενη λύση η οποία θυσιάζει λίγη από την ακρίβεια έναντι του υπολογιστικού και επικοινωνιακού κόστους. Η διακριτοποίηση μεγάλων συνόλων δεδομένων με την τεχνική Απεικόνισης/Μείωσης γενικότερα δεν είναι εύκολη διαδικασία και κατά συνέπεια χρειάζεται περεταίρω έρευνα προς αυτή την κατεύθυνση από την οποία θα μπορούσαν να προκύψουν αποδοτικές και κλιμακούμενες μέθοδοι διακριτοποίησης. Άλλο ένα σημαντικό θέμα που πρέπει να διερευνηθεί είναι η απόδοση του αλγόριθμου στην αναζήτηση του χώρου κανόνων με σύνολα δεδομένων μεγάλων (10^3) ή πολύ μεγάλων (10^6) διαστάσεων.

Σε αυτή την ενότητα, έγιναν εκτενείς αξιολογήσεις της προτεινόμενης προσέγγισης από τρεις προοπτικές: την ακρίβεια του μοντέλου κατηγοριοποίησης, το επικοινωνιακό κόστος και την

παράλληλη επίδοση. Τα αποτελέσματα δείχνουν ότι ο αλγόριθμος πετυχαίνει ακρίβεια συγκρίσιμη με αυτή που πετυχαίνουν γνωστοί αλγόριθμοι επαγωγής κανόνων και ότι κλιμακώνεται με την αύξηση των παραδειγμάτων. Κατά συνέπεια μπορεί να φανεί χρήσιμος για ένα μεγάλο εύρος εφαρμογών.

Κεφάλαιο 3

Δομική Ανάλυση διεπαφών αναζήτησης σε μεγάλη κλίμακα και λειτουργική κατηγοριοποίησή τους

3.1 Εισαγωγή

Βάσει μελετών, ο Παγκόσμιος Ιστός έχει αναγνωριστεί ότι αποτελείται από δυο μέρη: α) τον Επιφανειακό Ιστό του οποίου το περιεχόμενο μπορεί εύκολα να προσπελαστεί και να ενταχθεί σε ευρετήρια για να είναι διαθέσιμο προς αναζήτηση και β) τον Κρυμμένο Ιστό του οποίου το περιεχόμενο, παρόλο που είναι δημόσια διαθέσιμο, δεν μπορεί να προσπελαστεί από τις αυτοματοποιημένες μηχανές προσκομιδής περιεχομένου επειδή δημιουργείται δυναμικά σαν αποτέλεσμα της υποβολής ερωτημάτων σε φόρμες αναζήτησης. Έτσι το περιεχόμενο του Κρυμμένου Ιστού παραμένει μη διαθέσιμο στις μηχανές αναζήτησης και κατ' επέκταση στον τελικό χρήστη.

Στο κεφάλαιο αυτό μετά από τον ορισμό και την περιγραφή των κυριότερων χαρακτηριστικών του Κρυμμένου Ιστού, γίνεται μια επισκόπηση των προσεγγίσεων που έχουν προταθεί στη βιβλιογραφία για την προσπέλαση του περιεχομένου του. Το πρώτο βήμα σε κάθε μια από αυτές τις προσεγγίσεις είναι η αναγνώριση των διεπαφών ως προς τη λειτουργία τους. Με άλλα λόγια, οι προσεγγίσεις αυτές χρειάζονται σαν είσοδο, διεπαφές που όντως προορίζονται για αναζήτηση και όχι για άλλο σκοπό, όπως την εγγραφή ή τη σύνδεση σε έναν ιστότοπο, την υποβολή κάποιου σχολίου κλπ.

Η αναγνώριση των διεπαφών αναζήτησης τους αντιμετωπίζεται σαν πρόβλημα κατηγοριοποίησης και για την επίλυσή του χρησιμοποιείται ο αλγόριθμος επαγωγής κανόνων που παρουσιάστηκε στο Κεφάλαιο 2. Για την εκπαίδευση του κατηγοριοποιητή, επιλέχθηκε ένα δείγμα από το σύνολο των διεπαφών που εξήχθησαν από το σύνολο δεδομένων Yahoo L11. Μετά από ανάλυση του συγκεκριμένου συνόλου, επιλέγονται οι ιδιότητες του συνόλου εκπαίδευσης, με κύριο στόχο να είναι ανεξάρτητες της γλώσσας στην οποία είναι γραμμένες, έτσι ώστε να μπορεί να εφαρμοστεί σε όλον τον Παγκόσμιο Ιστό. Τέλος μετά από μια εμπειρική επιλογή των πιο κατάλληλων παραμέτρων για την προσέγγιση, γίνεται μια πειραματική μελέτη για την αξιολόγηση της επίδοσής της και μια σύγκριση με άλλους αλγόριθμους κατηγοριοποίησης.

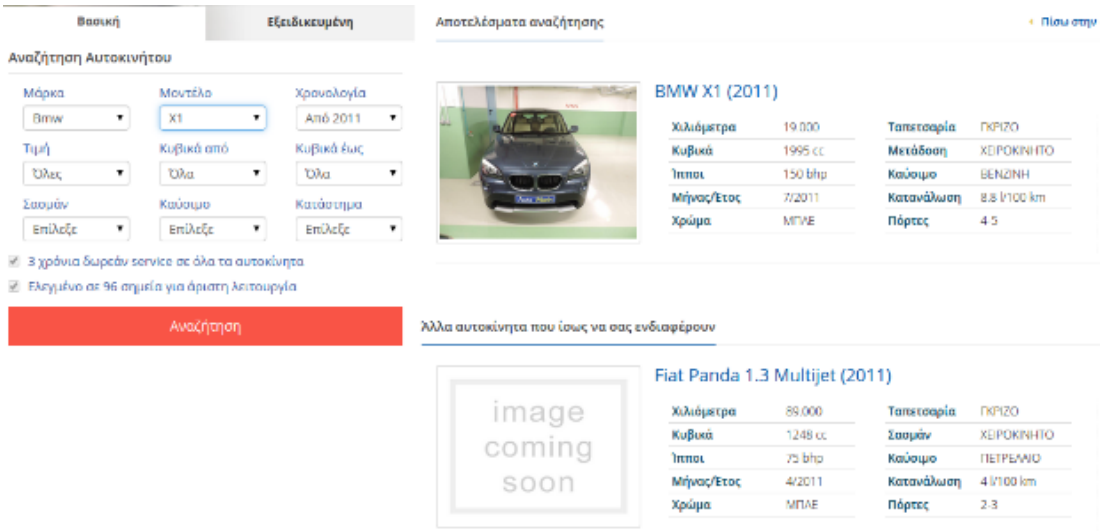
3.2 Ο Κρυμμένος και ο Επιφανειακός Παγκόσμιος Ιστός

Ο Παγκόσμιος Ιστός (World Wide Web) είναι το μεγαλύτερο πληροφοριακό σύστημα και η μεγαλύτερη πηγή δεδομένων στον κόσμο. Ο Ιστός αποτελεί πλέον αναπόσπαστο κομμάτι της καθημερινότητας εκατομμυρίων ανθρώπων [71], που απολαμβάνουν μέσω αυτού μια πληθώρα υπηρεσιών, όπως την ενημέρωση, την ψυχαγωγία, τα κοινωνικά μέσα και τα ηλεκτρονικά καταστήματα. Ο Ιστός παρουσιάζει πολλαπλά ιδιαίτερα χαρακτηριστικά που καθιστούν την εύρεση πληροφοριών σε αυτόν ένα ιδιαίτερα σύνθετο εγχείρημα. Παρόλο που το ακριβές μέγεθός του δεν μπορεί να μετρηθεί με ακρίβεια, ο Ιστός μπορεί να μελετηθεί και να κατανοηθεί, βάσει των υπόλοιπων ιδιοτήτων του όπως το δυναμισμό του περιεχομένου του και τη διασύνδεση των πληροφοριακών μονάδων του.

Ο τεράστιος όγκος δεδομένων που υπάρχει στον Ιστό, έχει καταστήσει αδύνατη την εύρεση πληροφοριών με χειροκίνητο τρόπο. Η μόνη φιλική προς το χρήστη μέθοδος για την ικανοποίηση των πληροφοριακών αναγκών του, είναι η χρήση αυτοματοποιημένων μηχανών αναζήτησης. Οι μηχανές αναζήτησης στον Ιστό αποτελούν συστήματα λογισμικού σχεδιασμένα για την αναζήτηση πληροφοριών από διάφορες, συχνά ετερογενείς πηγές. Ανάλογα με το εύρος και το είδος των πληροφοριακών πηγών που καλύπτουν, οι μηχανές αναζήτησης διακρίνονται σε α) γενικού σκοπού ή οριζόντιες, β) ειδικού σκοπού ή κάθετες, γ) μετα-αναζήτησης και δ) βάσεων δεδομένων. Οι μηχανές αναζήτησης γενικού σκοπού (general purpose search engines) βασίζονται σε ευρετήρια που κατασκευάζονται και ενημερώνονται αυτόματα από δεδομένα που συλλέγουν ειδικά προγράμματα προσκομιδής περιεχομένου (web crawlers) από όλο το εύρος του Παγκόσμιου Ιστού. Οι κάθετες μηχανές αναζήτησης (vertical search engines) αποτελούν ειδική περίπτωση της πρώτης κατηγορίας, αφού εστιάζουν σε κάποιο συγκεκριμένο κομμάτι του Ιστού, όπως μια θεματική κατηγορία ή κάποιο είδος περιεχομένων. Οι μηχανές μετα-αναζήτησης (meta-search engines) είναι εργαλεία που παράγουν τα αποτελέσματά τους από το συνδυασμό των αποτελεσμάτων άλλων μηχανών αναζήτησης. Τέλος, οι μηχανές αναζήτησης βάσεων δεδομένων (database search engines), περιορίζονται στην εύρεση πληροφοριών από τη βάση δεδομένων ενός συγκεκριμένου ιστότοπου και συναντώνται συχνά σε ηλεκτρονικά καταστήματα. Οι ιστότοποι που βασίζονται σε βάσεις δεδομένων και παρέχουν πρόσβαση σε αυτά μόνο μέσα από μια μηχανή αναζήτησης καλούνται συχνά και ως βάσεις δεδομένων Ιστού (web databases).

Κάθε μηχανή αναζήτησης στον Ιστό διαθέτει τουλάχιστον μια φόρμα αναζήτησης, μέσω τις οποίας επιτρέπεται η υποβολή συγκεκριμένων ερωτημάτων, βάσει των οποίων σχηματίζονται οι σελίδες αποτελεσμάτων σαν απάντηση στα ερωτήματα αυτά. Οι φόρμες αναζήτησης (search forms), που είναι επίσης γνωστές στη βιβλιογραφία και ως διεπαφές ερωτημάτων (query interfaces) ή διεπαφές αναζήτησης (search interfaces) αποτελούνται από ένα σύνολο πεδίων (input fields) τα οποία διαμορφώνουν τις συνθήκες των ερωτημάτων (query conditions) που υποβάλλονται στη μηχανή αναζήτησης. Οι διεπαφές αναζήτησης βάσεων δεδομένων είναι γενικά πιο σύνθετες από τις

τυπικές διεπαφές των μηχανών αναζήτησης γενικού σκοπού. Αυτό οφείλεται στο γεγονός ότι οι πρώτες μπορεί να απαιτούν τον προσδιορισμό πολλαπλών συνθηκών (π.χ. την ημερομηνία άφιξης και την ημερομηνία αναχώρησης) ενώ οι δεύτερες χρειάζονται απλώς μία λέξη-κλειδί ή μια φράση. Στην Εικόνα 3.1 φαίνεται ένα παράδειγμα σύνθετης διεπαφής ερωτημάτων για ένα κατάστημα αυτοκινήτων[1]. Η διεπαφή αυτή περιέχει πολλαπλά πεδία αναζήτησης που εξειδικεύουν το ερώτημα προς υποβολή. Τα αποτελέσματα που επιστρέφονται απαντούν ακριβώς στο ερώτημα που εισήχθη. Στην Εικόνα 3.2 φαίνεται ένα παράδειγμα μηχανής γενικού σκοπού, όπου είναι δυνατόν να εισαχθεί μια αυθαίρετη φράση. Τα αποτελέσματα είναι μεν σχετικά με το ερώτημα αλλά είναι και γενικά.



Εικόνα 3.1 – Η διεπαφή αναζήτησης ενός καταστήματος αυτοκινήτων.

Ένα κύριο χαρακτηριστικό των βάσεων Ιστού, είναι ότι οι σελίδες με το περιεχόμενό τους παράγονται δυναμικά, σαν αποτέλεσμα της υποβολής ενός ερωτήματος μέσω της διεπαφής ερωτημάτων τους. Επίσης, τόσο οι βάσεις αυτές όσο και τρίτοι ιστότοποι, συχνά παρέχουν ελάχιστους, ή και καθόλου συνδέσμους προς τις σελίδες αυτές. Αυτό το φαινόμενο αποτελεί σημαντικό εμπόδιο για τη διαδικασία προσκομιδής σελίδων από τις μηχανές αναζήτησης γενικού σκοπού, αφού αυτή βασίζεται στη διασύνδεση των σελίδων για την άφιξη σε μια σελίδα και τη μεταφόρτωσή της. Η διαδικασία αυτή σταματά στις διεπαφές ερωτημάτων και κατά συνέπεια ένας μεγάλος όγκος περιεχομένου χάνεται και παραμένει αδιάθετος στον τελικό χρήστη της μηχανής αναζήτησης. Εξαιτίας του φαινομένου αυτού ο Παγκόσμιος Ιστός έχει χωριστεί σε δυο μέρη: α) τον Επιφανειακό Ιστό (Surface Web) τον οποίο τα προγράμματα προσκομιδής σελίδων μπορούν να προσπελάσουν χωρίς προβλήματα και β) τον Κρυμμένο Ιστό (Hidden Web) γνωστό και ως Βαθύ Ιστό (Deep Web) που αποτελείται από σελίδες που παράγονται δυναμικά σαν αποτέλεσμα σε

[1] http://www.automarin.gr/

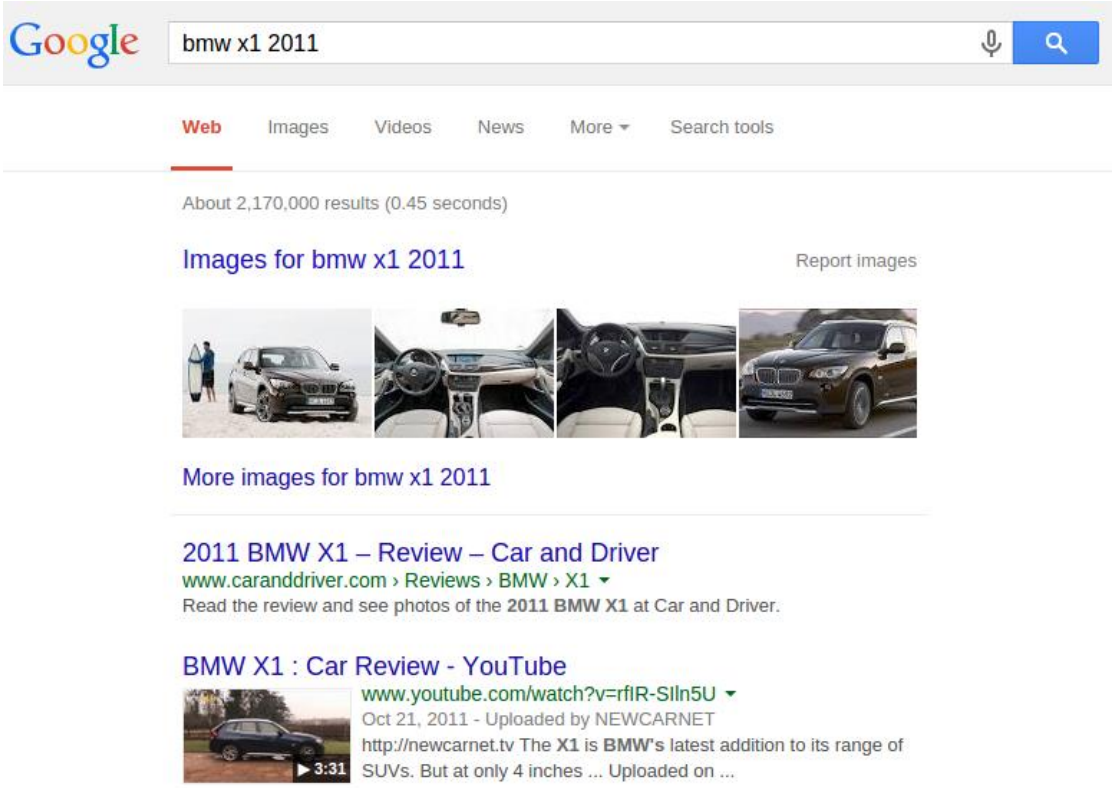
ερωτήματα που υποβάλλονται μέσα από διεπαφές αναζήτησης. Ο Κρυμμένος Ιστός δεν θα πρέπει να συγχέεται με τον Σκοτεινό Ιστό (Dark Web) που απλώς αποτελεί ένα μικρό υποσύνολό του. Ο Σκοτεινός Ιστός αποτελείται κατά κύριο λόγο από περιεχόμενο η πρόσβαση στο οποίο γίνεται μόνο μέσα από ειδικό λογισμικό για την είσοδο σε εικονικά ιδιωτικά δίκτυα (virtual private networks) όπως για παράδειγμα το TOR [72].

3.3 Χαρακτηριστικά του Κρυμμένου Ιστού

Ο Κρυμμένος Ιστός έχει αναγνωριστεί ότι αποτελείται από περιεχόμενο τάξεις μεγέθους περισσότερο από το αντίστοιχο του Επιφανειακού Ιστού, με ποιοτικές πληροφορίες, η αξία των οποίων προέρχεται κυρίως από τη δομή των δεδομένων του. Περιέχει εκατομμύρια σελίδες, που προστίθενται, αφαιρούνται και ενημερώνονται δυναμικά και που μπορούν να προσπελαστούν μόνο μέσα από διεπαφές ερωτημάτων που είναι διάσπαρτες στον Επιφανειακό Ιστό. Σύμφωνα με μια μελέτη του 2007 [73], υπάρχουν ήδη 25 εκατομμύρια πηγές δεδομένων από τον κρυμμένο ιστό. Το περιεχόμενο του Κρυμμένου Ιστού παρουσιάζεται συχνά στους χρήστες ως εγγραφές σε πινακοειδή μορφή μέσα σε δυναμικά παραγόμενες σελίδες, ως αποτέλεσμα της υποβολής κάποιου ερωτήματος σε μια διεπαφή ερωτημάτων Οι εγγραφές αυτές μπορεί να ακολουθούν κάποιο σχήμα με ονομασμένες στήλες συγκεκριμένων τύπων δεδομένων και κατά συνέπεια κάθε εγγραφή μπορεί να θεωρηθεί άμεσα ή έμμεσα μια εγγραφή βάσης δεδομένων. Δυστυχώς τέτοια σχήματα μπορεί να μην είναι γνωστά εκ των προτέρων και κατά συνέπεια θα πρέπει να καταβληθεί σημαντική προσπάθεια για την εξαγωγή τους.

Από την προοπτική του χρήστη, το περιεχόμενο του Κρυμμένου Ιστού μπορεί να είναι αρκετές φορές προτιμότερο, επειδή συνήθως ανήκει σε μια συγκεκριμένη θεματική κατηγορία και μπορεί να προσπελαστεί μόνο μέσα από διεπαφές που επιτρέπουν την υποβολή αυστηρών ερωτημάτων, σε αντίθεση με τις χαλαρές φράσεις ή λέξεις-κλειδιά που υποβάλλονται στις μηχανές αναζήτησης γενικού σκοπού. Ωστόσο, όταν χρειάζεται ο συνδυασμός πληροφοριών από πολλαπλές πηγές, όχι μόνο θα πρέπει να γνωρίζει πού θα βρει αυτές τις πηγές αλλά θα χρειαστεί να υποβάλλει το ίδιο ερώτημα πολλαπλές φορές.

Πέραν του μεγέθους του, ο Κρυμμένος Ιστός καλύπτει μια ευρεία γκάμα θεματικών κατηγοριών ή αντικειμένων η οποία αυξάνεται συνεχώς. Επίσης η πλειονότητα των σελίδων που βρίσκονται σε αυτόν είναι δημόσια προσπελάσιμες, δηλαδή δεν χρειάζεται κάποια εγγραφή και αυθεντικοποίηση του χρήστη για την προσπέλασή τους. Η μεγάλη πλειοψηφία των διεπαφών ερωτημάτων επίσης, μπορούν να εντοπιστούν ακολουθώντας το πολύ τρεις συνδέσμους από την κεντρική σελίδα του ιστότοπου στον οποίο ανήκουν [74] [75]. Όλες οι παραπάνω ιδιότητες καθιστούν προφανείς τις δυσκολίες εύρεσης πληροφοριών στον Κρυμμένο Ιστό, αλλά και την αναγκαιότητα της δημιουργίας αυτοματοποιημένων προσεγγίσεων που θα μπορούσαν να τις εκμεταλλευτούν προς όφελός τους.



Εικόνα 3.2 - Η διεπαφή μιας μηχανής αναζήτησης γενικού σκοπού

Έχουν αναπτυχθεί διάφορες μέθοδοι προς επίτευξη αυτού του στόχου. Οι περισσότερες από αυτές μπορούν να ενταχθούν σε δυο ευρείες μεθοδολογίες. Μια από αυτές περιλαμβάνει τη λήψη εγγραφών από κάθε βάση δεδομένων μετά από αυτοματοποιημένη υποβολή ερωτημάτων, την τοποθέτησή τους σε ένα κεντρικό σύστημα και τη δημιουργία μιας νέας βάσης δεδομένων από τα δεδομένα αυτά. Μια άλλη προσέγγιση περιλαμβάνει την ενοποιημένη πρόσβαση σε πολλαπλές υπάρχουσες βάσεις δεδομένων, συνήθως της ίδιας θεματικής κατηγορίας, χωρίς να έχουν προσκομιστεί τα περιεχόμενά τους εκ των προτέρων. Και οι δυο προαναφερθείσες μεθοδολογίες απαιτούν αλληλεπίδραση με τη διεπαφή αναζήτησης κάθε βάσης δεδομένων, ενώ η δεύτερη χρειάζεται επιπρόσθετα και την ενοποίηση των διεπαφών αναζήτησης που ανήκουν στην ίδια θεματική κατηγορία. Στην επόμενη παράγραφο αναλύονται οι δυο προσεγγίσεις διεξοδικότερα.

3.4 Μέθοδοι πρόσβασης σε δεδομένα του Κρυμμένου Ιστού

Υπάρχουν δυο γενικές μέθοδοι που επιτρέπουν την πρόσβαση σε δεδομένα στον κρυμμένο ιστό. Η πρώτη είναι γνωστή ως ανάδυση (surfacing) και η δεύτερη είναι γνωστή ως μετα-αναζήτηση (meta-search). Στην πρώτη μέθοδο, αρχικά τα δεδομένα συλλέγονται από τις πηγές και στη συνέχεια κατασκευάζεται ένα ευρετήριο αναζήτησης για τα δεδομένα αυτά, το οποίο χρησιμοποιείται για να παρέχεται ενοποιημένη πρόσβαση στους χρήστες. Στη μέθοδο της μετα-

αναζήτησης, δημιουργείται ένα σύστημα αναζήτησης το οποίο μέσω μιας ενοποιημένης διεπαφής ερωτημάτων προωθεί τα ερωτήματα των χρηστών σε άλλες βάσεις δεδομένων, λαμβάνει τα αποτελέσματα από κάθε μια ξεχωριστά και να τα παρουσιάζει στον τελικό χρήστη μετά από κάποια κατάταξη. Κάθε μια από τις προσεγγίσεις έχει τα δικά της πλεονεκτήματα και μειονεκτήματα. Ακολουθεί η συνοπτική περιγραφή τους στις επόμενες παραγράφους.

3.4.1 Ανάδυση

Η ανάδυση δεδομένων είναι μια μέθοδος που προτιμάται από τις μεγάλες μηχανές αναζήτησης γενικού σκοπού, αφού μπορεί να εκμεταλλευτεί στο έπακρο την υπάρχουσα υποδομή τους. Υπάρχουν δυο προσεγγίσεις ανάδυσης δεδομένων που διαφέρουν μόνο στον τρόπο προσκομιδής των δεδομένων από τον Κρυμμένο Ιστό. Στην πρώτη προσέγγιση που είναι γνωστή και ως τροφοδοσία δεδομένων (data feed), οι ίδιοι ιδιοκτήτες των δεδομένων αναλαμβάνουν να αποστείλουν τα δεδομένα τους μέσω ειδικών υπηρεσιών στις μηχανές αναζήτησης για να εισαχθούν στα ευρετήριά τους και να καταστούν υποψήφια αποτελέσματα σε πιθανές ερωτήσεις των χρηστών. Τα δεδομένα που αποστέλλονται κατ' αυτόν τον τρόπο είναι συνήθως προϊόντα [76] [77]. Ωστόσο με αυτόν τον τρόπο μόνο ένα μικρό ποσοστό του Κρυμμένου Ιστού καθίσταται γνωστό στις μηχανές αναζήτησης, καθώς ελάχιστοι ιδιοκτήτες αναρτούν οικειοθελώς τα δεδομένα τους στις υπηρεσίες αυτές.

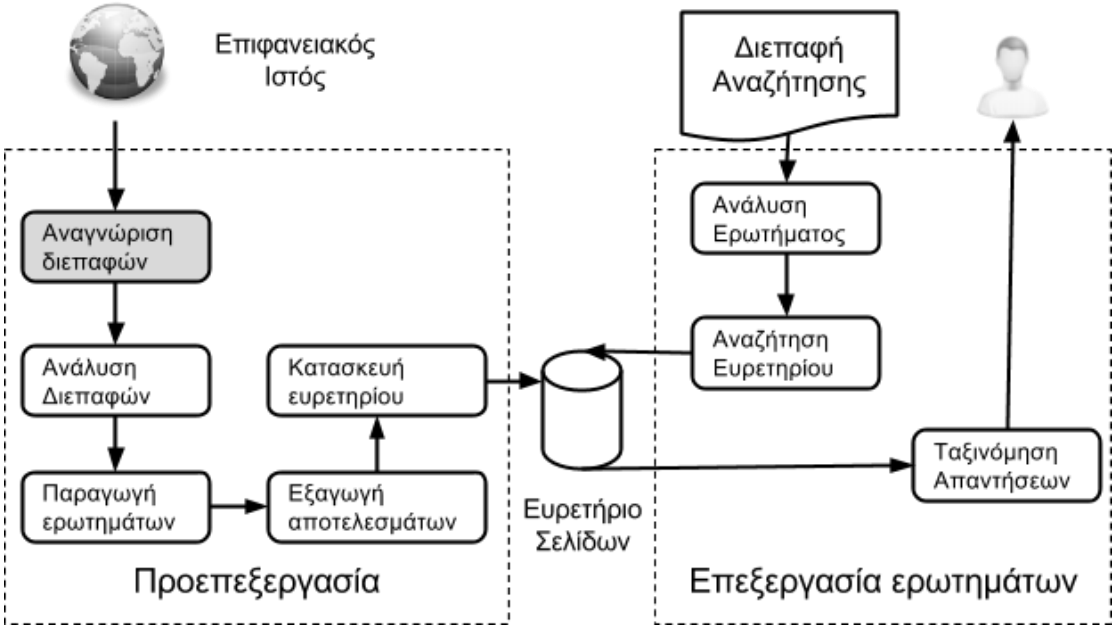
Στη δεύτερη προσέγγιση που καλείται προσκομιδή δεδομένων από τον Κρυμμένο Ιστό (deep web crawling), το περιεχόμενο του Κρυμμένου Ιστού προσκομίζεται στο κεντρικό σύστημα της μηχανής αναζήτησης μετά από αυτοματοποιημένη υποβολή ερωτημάτων στις διεπαφές αναζήτησης των κατά τόπους βάσεων. Μέχρι να καταστούν διαθέσιμα τα δεδομένα του Κρυμμένου Ιστού στον τελικό χρήστη, μεσολαβούν μια σειρά βημάτων που συνοψίζονται στην Εικόνα 3.3 και παρουσιάζονται αναλυτικά στις επόμενες παραγράφους.

3.4.1.1 Αυτόματη αναγνώριση των διεπαφών αναζήτησης

Οι διεπαφές ερωτημάτων βρίσκονται στον Επιφανειακό Ιστό, κάτι που σημαίνει ότι είναι δυνατή η εύρεσή τους με ένα συμβατικό πρόγραμμα προσκομιδής ιστοσελίδων. Η εύρεση αυτή θεωρείται απλή για τις κυριότερες μηχανές αναζήτησης γενικού σκοπού [78], αφού ήδη διαθέτουν ένα αντιπροσωπευτικό δείγμα του Επιφανειακού Ιστού το οποίο με τη σειρά του θα περιέχει έναν σημαντικό αριθμό διεπαφών ερωτημάτων. Η όλη διαδικασία όμως μπορεί να συνοψιστεί στα εξής βήματα: α) λήψη σελίδων από τον Επιφανειακό Ιστό με ένα συμβατικό πρόγραμμα προσκομιδής σελίδων, β) έλεγχος κάθε σελίδας για την ύπαρξη διεπαφών και γ) αναγνώριση της λειτουργικότητας της διεπαφής. Η δυσκολία της όλης διαδικασίας βρίσκεται στο τρίτο βήμα όπου χρειάζεται να διαχωρίζονται οι διεπαφές αναζήτησης από τις διεπαφές που έχουν άλλη λειτουργία όπως τη δημοσίευση σχολίων, την εγγραφή ή τη σύνδεση με έναν ιστότοπο. Το πρόβλημα αυτό αποτελεί το κύριο θέμα του κεφαλαίου αυτού και αναλύεται διεξοδικότερα στην παράγραφο 3.5.

3.4.1.2 Ανάλυση διεπαφών ερωτημάτων

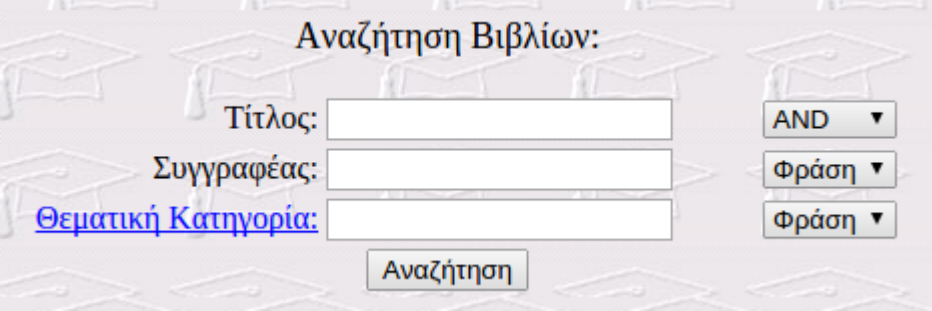
Η ανάλυση των διεπαφών είναι μια διαδικασία γνωστή και ως κατανόηση των φορμών (form understanding) [79] και αποσκοπεί στην επίτευξη τριών στόχων:



Εικόνα 3.3 - Η Μεθοδολογία της ανάδυσης δεδομένων από τον Κρυμμένο Ιστό

- *την αναγνώριση της υποκείμενης βάσης δεδομένων*, όπου αφενός αναγνωρίζεται η προέλευση των δεδομένων μέσω του ονόματος και της εφαρμογής του εξυπηρέτη όπου υποβάλλεται το ερώτημα και αφετέρου αναγνωρίζονται τα πεδία του ερωτήματος και οι προεπιλεγμένες τιμές τους (αν αυτές παρέχονται).
- *την κατανόηση της σημασίας κάθε πεδίου εισόδου*, όπου εξάγεται η ετικέτα κάθε πεδίου [80] και προσδιορίζονται οι κατάλληλες τιμές που μπορούν να εισαχθούν στο πεδίο με ιδιαίτερη έμφαση στα κειμενικά πεδία, όπου οι τιμές που μπορούν να εισαχθούν είναι ουσιαστικά άπειρες.
- *τον προσδιορισμό του συνόλου έγκυρων ερωτημάτων για μια διεπαφή*, όπου κατασκευάζονται ερωτήματα που είναι αποδεκτά από τη διεπαφή ανεξάρτητα από το αν αυτή επιστρέφει αποτελέσματα μετά από την υποβολή τους ή όχι. Για το χαρακτηρισμό ενός ερωτήματος ως έγκυρο, εισήχθη η έννοια των ατομικών ερωτημάτων (atomic query) [81] τα οποία είναι τα ελάχιστα σύνολα πεδίων αναζήτησης, οι τιμές των οποίων πρέπει να παρέχονται. Για παράδειγμα καθένα από τα τρία πεδία της Εικόνα 3.4, αποτελεί ένα ατομικό ερώτημα. Η αναγνώριση ατομικών ερωτημάτων γίνεται μόνο με μεθόδους που περιλαμβάνουν την υποβολή ερωτημάτων και την ανάλυση των αποτελεσμάτων τους και βοηθάει αρκετά στην απαλοιφή της πιθανότητας παραγωγής άκυρων ερωτημάτων,

βελτιώνοντας έτσι την αποτελεσματικότητα της προσκομιδής δεδομένων από τον Κρυμμένο Ιστό.



Εικόνα 3.4- Η διεπαφή ερωτημάτων για την αναζήτηση βιβλίων της βιβλιοθήκης της σχολής ΗΜΜΥ του ΕΜΠ

3.4.1.3 Αυτόματη κατασκευή ερωτημάτων

Η διεπαφή ερωτημάτων μιας βάσης ιστού μπορεί να έχει πολλαπλά πεδία εισόδου και το πλήθος όλων των πιθανών ερωτημάτων που μπορούν να υποβληθούν σε αυτό υπολογίζεται με το καρτεσιανό γινόμενο των πιθανών τιμών κάθε πεδίου του. Είναι προφανές πως η εξαντλητική κατασκευή όλων των πιθανών ερωτημάτων με τις τιμές των πεδίων δεν είναι πρακτικά εφαρμόσιμη, ειδικά όταν περιέχονται και κειμενικά πεδία, αφού χρειάζονται σημαντικοί πόροι τόσο στην πλευρά του κατασκευαστή των ερωτημάτων όσο και στην πλευρά της διεπαφής. Αντίθετα ο στόχος στην κατασκευή ερωτημάτων για υποβολή σε διεπαφές αναζήτησης και την εξαγωγή των αποτελεσμάτων τους πρέπει να είναι η μεγιστοποίηση του πλήθους των αποτελεσμάτων αυτών με το ελάχιστο δυνατό πλήθος ερωτημάτων. Με στόχο την αυτοματοποίηση της κατασκευής ερωτημάτων, έχουν προταθεί διάφορες προσεγγίσεις στη βιβλιογραφία, στο επίκεντρο των οποίων βρίσκονται τα ερωτήματα που περιλαμβάνουν κειμενικά πεδία εισόδου.

Η μέθοδος που προτείνεται στην εργασία [82] αποτελεί την προσέγγιση της Google για τη λήψη δεδομένων από τον Κρυμμένο Ιστό. Χρησιμοποιεί τις φόρμες που έχουν ήδη εντοπιστεί από τους προσκομιστές δεδομένων της Google για τον Επιφανειακό Ιστό. Η μέθοδος αυτή εισάγει την έννοια των ενημερωτικών προτύπων ερωτημάτων (informative query templates), που αποτελούν ένα υποσύνολο των πεδίων εισόδου σε μια διεπαφή ερωτημάτων που μπορούν να παράξουν επαρκώς διαφορετικά αποτελέσματα για διαφορετικές τιμές στα πεδία εισόδου τους. Η μέθοδος αναγνωρίζει τις κατάλληλες τιμές για κάθε πεδίο κειμένου ακολουθώντας την τεχνική της επαναληπτικής υποβολής ερωτημάτων. Ξεκινά με μερικές αρχικές λέξεις που έχουν εξαχθεί από τη σελίδα με τη διεπαφή ερωτημάτων, υποβάλλει κάθε λέξη στο πεδίο κειμένου, αναγνωρίζει νέες λέξεις από τα αποτελέσματα, ομαδοποιεί τις νέες λέξεις, έτσι ώστε οι λέξεις στην ίδια ομάδα να είναι συναφείς εννοιολογικά (συνώνυμες) και επιλέγει μια λέξη από κάθε ομάδα για την επόμενη επανάληψη. Η μέθοδος αυτή επικεντρώνεται λιγότερο στην εξαντλητική προσκομιδή δεδομένων από κάθε βάση, και περισσότερο στην επίτευξη καλής κάλυψης από ένα μεγάλο πλήθος βάσεων με περιορισμένες

υποβολές. Ωστόσο παρουσιάζει κάποιους περιορισμούς, όπως για παράδειγμα το γεγονός ότι αλληλεπιδρά μόνο με φόρμες τύπου GET, κάτι που οδηγεί στο συμπέρασμα ότι με τη μέθοδο αυτή χάνονται αρκετά δεδομένα, αφού όπως φαίνεται από τα αποτελέσματα της στατιστικής μελέτης του συνόλου δεδομένων Yahoo L11 της παραγράφου 3.6, το πλήθος των διεπαφών αναζήτησης τύπου POST είναι σημαντικό.

3.4.1.4 Αυτόματη εξαγωγή αποτελεσμάτων

Η τρίτη και τελευταία διαδικασία της ανάδυσης δεδομένων από τον Κρυμμένο Ιστό, περιλαμβάνει την εξαγωγή δεδομένων από τα αποτελέσματα που επιστρέφονται από την προηγούμενη φάση. Αυτή η διαδικασία είναι αναγκαία για την επίτευξη του απώτερου στόχου της ανάδυσης που είναι η επεξεργασία και η τοποθέτηση στο ευρετήριο των μηχανών αναζήτησης των δεδομένων που είναι κρυμμένα πίσω από τις διεπαφές. Στα πλαίσια της διαδικασίας αυτής, έχουν προταθεί διάφορες προσεγγίσεις που περιλαμβάνουν α) την εξαγωγή των αποτελεσμάτων από τις σελίδες που επιστρέφονται σαν αποτέλεσμα σε ένα ερώτημα και β) την ανάθεση σημασιολογικών ετικετών στα δεδομένα κάθε εγγραφής. Ένα τμήμα της σελίδας με τα αποτελέσματα ενός ερωτήματος στη βάση του ηλεκτρονικού καταστήματος αυτοκινήτων automarin.gr φαίνεται στην Εικόνα 3.1. Μια τέτοια βάση επιστρέφει δυναμικά παραχθείσες σελίδες που περιέχουν τις εγγραφές που είναι σχετικές με το υποβληθέν ερώτημα, και συνδέσμους στις επόμενες σελίδες αποτελεσμάτων, αν τα αποτελέσματα είναι περισσότερα απ' όσα χωράει μια σελίδα, ενώ συχνά περιέχουν περιττές πληροφορίες όπως για παράδειγμα διαφημίσεις.

Η αυτόματη εξαγωγή αποτελεσμάτων αναζητήσεων είναι μια ειδική περίπτωση της εξαγωγής πληροφοριών από τον Ιστό (web information extration) [83]. Αυτές οι τεχνικές περιλαμβάνουν α) τις τεχνικές που χρησιμοποιούν μόνο τη σήμανση (δηλαδή τον κώδικα HTML) των σελίδων απόκρισης [84] β) τις τεχνικές που πέραν της σήμανσης της HTML, λαμβάνουν υπόψη και την οπτική πληροφορία των σελίδων αποτελεσμάτων [85], και τέλος γ) αυτές που ενσωματώνουν γνώση για τη θεματική κατηγορία όπου ανήκει η βάση δεδομένων [86]. Ένα κοινό χαρακτηριστικό που μοιράζονται αρκετές από τις τεχνικές αυτές είναι ότι κατά τη διάρκεια της εξαγωγής δεδομένων από τις σελίδες αποτελεσμάτων, γίνεται προσπάθεια να αναγνωριστεί το τμήμα με τις απαντήσεις και να εξαχθούν οι εγγραφές χρησιμοποιώντας διαχωριστές που έχουν προσδιοριστεί αυτόματα κατά τη διαδικασία.

3.4.1.5 Πλεονεκτήματα και μειονεκτήματα

Τα πλεονεκτήματα της μεθόδου ανάδυσης δεδομένων περιλαμβάνουν:

- τη δυνατότητα προεπεξεργασίας των δεδομένων, αφού αυτά είναι στην κατοχή του συστήματος (π.χ. η σύνδεση δεδομένων που προέρχονται από διαφορετικές πηγές)
- τη δημιουργία ευρετηρίου αναζήτησης για τη γρήγορη επεξεργασία των ερωτημάτων.

- τα δεδομένα που συλλέγονται από τον κρυμμένο ιστό μπορούν να αναζητηθούν μαζί με τα δεδομένα που συλλέγονται από τον επιφανειακό ιστό με τον ίδιο τρόπο, όπως γίνεται στη Google [82] [87].

Από την άλλη, η προσέγγιση αυτή έχει μερικούς περιορισμούς:

- ένα μεγάλο κομμάτι του κρυμμένου ιστού παραμένει κρυμμένο, αφού οι υπάρχουσες τεχνικές αδυνατούν να το προσπελάσουν.
- είναι εξαιρετικά δύσκολη η συντήρηση της φρεσκάδας των αναδυθέντων δεδομένων, αφού τα προγράμματα προσκόμισης δεν μπορούν να συμβαδίσουν με το δυναμισμό των δεδομένων του κρυμμένου ιστού.

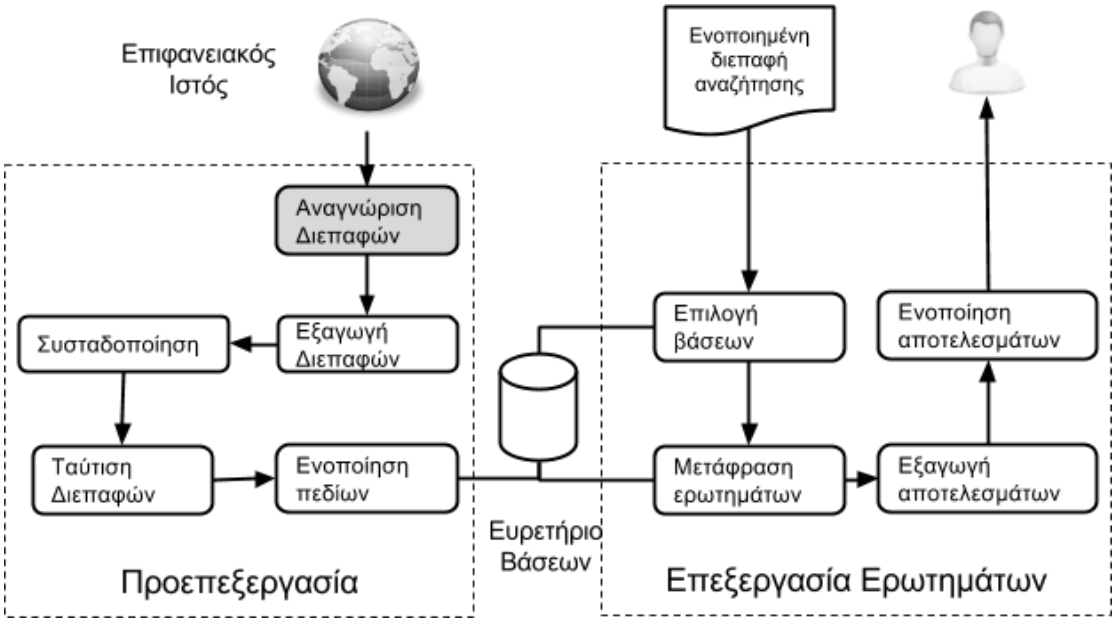
3.4.2 Μέτα-Αναζήτηση

Στόχος της μεθόδου της μετα-αναζήτησης, είναι η δημιουργία ενός συστήματος ενοποίησης πολλαπλών Βάσεων Ιστού. Τέτοια συστήματα παρουσιάζουν αρκετές ομοιότητες με τις μηχανές μετα-αναζήτησης γενικού σκοπού (meta-search engines) ή τα ομόσπονδα συστήματα αναζήτησης (federated search systems). Τα συστήματα μετα-αναζήτησης παρέχουν τη δική τους ενοποιημένη διεπαφή ερωτημάτων και τα ερωτήματα που υποβάλλονται σε αυτή αρχικά μεταφράζονται σε μια μορφή που είναι αποδεκτή από τις αντίστοιχες διεπαφές όπου και τελικά προωθούνται. Όταν το σύστημα ενοποίησης λαμβάνει τα αποτελέσματα από τις κατά τόπους βάσεις, τα ενοποιεί και τα εμφανίζει στον τελικό χρήστη. Η δημιουργία τέτοιων συστημάτων, περιέχει αρκετές προκλήσεις καθώς οι βάσεις είναι ποικιλόμορφες (περιέχουν δεδομένα διαφορετικά δομημένα) και συνήθως ανεξάρτητες (κάθε μια σχεδιάζεται και λειτουργεί διαφορετικά). Οι τεράστιες διαφορές που υπάρχουν στις βάσεις δεδομένων καθιστούν αδύνατο να κατασκευαστεί ένα μοναδικό κεντρικό σύστημα ενοποίησης που να τις καλύπτει όλες και για το λόγο αυτό στην πράξη δημιουργούνται διαφορετικές μηχανές ενοποίησης για κάθε θεματική κατηγορία. Για παράδειγμα μπορεί να κατασκευαστεί ένα σύστημα ενοποίησης για όλες τις βάσεις δεδομένων στην κατηγορία των βιβλίων.

Η δημιουργία πολλαπλών συστημάτων ενοποίησης που καλύπτουν βάσεις διαφορετικών θεματικών περιοχών περιλαμβάνει σαν πρώτο βήμα την εύρεση των βάσεων. Στη συνέχεια χρειάζεται μια μέθοδος για την ομαδοποίηση των βάσεων δεδομένων σύμφωνα με τις θεματικές τους κατηγορίες, έτσι ώστε η κατασκευή των συστημάτων ενοποίησης να περιλαμβάνει μόνο βάσεις από κάθε κατηγορία.

Η διαδικασία της δημιουργίας μιας ενοποιημένης διεπαφής για κάθε θεματική περιοχή, μπορεί να θεωρηθεί ότι αποτελείται από πέντε κύρια βήματα όπως φαίνεται στην Εικόνα 3.5: α) Στο πρώτο βήμα εξάγεται το σχήμα από τον κώδικα HTML της διεπαφής ερωτημάτων κάθε βάσης. β) Στο δεύτερο βήμα, γίνεται η ομαδοποίηση των διεπαφών ως προς τη θεματική τους κατηγορία. γ) Στο τρίτο βήμα, αναγνωρίζονται τα πεδία που έχουν παρόμοια σημασία μεταξύ των διαφορετικών

διεπαφών αναζήτησης και αντιστοιχίζονται. δ) Στο τέταρτο βήμα, τα αντιστοιχισμένα πεδία των διαφορετικών διεπαφών, ενοποιούνται, συμπεριλαμβανομένων των διαφόρων χαρακτηριστικών, όπως των ονομάτων, των μορφών και των εξωτερικών τιμών των αντιστοιχισμένων ιδιοτήτων.



Εικόνα 3.5 - Η μεθοδολογία της μέτα-αναζήτησης στον Κρυμμένο Ιστό

Η επεξεργασία των ερωτημάτων που υποβάλλεται από τους χρήστες αποτελείται σε ανώτερο επίπεδο από τέσσερα βήματα: α) την επιλογή των βάσεων που είναι σχετικές με το ερώτημα, β) την μετάφραση των ερωτημάτων έτσι ώστε να είναι έγκυρα για τις διεπαφές κάθε μιας από τις βάσεις αυτές, γ) εξαγωγή των αποτελεσμάτων από τις βάσεις και δ) ενοποίηση των αποτελεσμάτων και παρουσίασή τους στο χρήστη.

Αρχικά, αφού ληφθεί ένα ερώτημα χρήστη από την ενοποιημένη διεπαφή, στο βήμα της επιλογής βάσεων, αποφασίζεται ποιές από τις υποστηριζόμενες βάσεις δεδομένων θα πρέπει να χρησιμοποιηθούν για να προωθηθεί το τρέχον ερώτημα [88]. Το βήμα αυτό είναι απαραίτητο ιδιαίτερα όταν το πλήθος των υποστηριζόμενων βάσεων είναι μεγάλο, σαν φιλτράρισμα των άσχετων βάσεων. Για μια σωστή επιλογή βάσεων, θα πρέπει να συγκεντρωθούν πληροφορίες για τα περιεχόμενα κάθε βάσης εκ των προτέρων.

Στο βήμα της μετάφρασης των ερωτημάτων, γίνεται αντιστοίχιση κάθε ερωτήματος που υποβάλλεται στη διεπαφή ερωτημάτων (τέτοια ερωτήματα καλούνται καθολικά ερωτήματα) σε ένα ή περισσότερα ερωτήματα (που είναι γνωστά ως τοπικά) τα οποία θα υποβάλλονται στην κάθε διεπαφή ερωτημάτων των επιλεγμένων βάσεων δεδομένων. Μερικές φορές μια καθολική διεπαφή μπορεί να μεταφράζεται σε πολλαπλά ερωτήματα για να υποβληθεί σε μια τοπική διεπαφή. Είναι επίσης πιθανόν μερικά καθολικά ερωτήματα να μην μπορούν να μεταφραστούν σε αντίστοιχα τοπικά, για λόγους ασυμφωνίας ευρών τιμών ή αναντιστοιχία ιδιοτήτων. Σε τέτοιες περιπτώσεις το καθολικό ερώτημα μεταφράζεται σε ένα πιο ευρύ τοπικό ερώτημα για να διασφαλιστεί η λήψη όλων των

αποτελεσμάτων που χρειάζονται. Ταυτόχρονα αφαιρούνται τα αποτελέσματα που δεν είναι σχετικά με το ερώτημα σύμφωνα με προκαθορισμένα κριτήρια.

Το βήμα της εξαγωγής αποτελεσμάτων είναι παρόμοιο με το αντίστοιχο βήμα της μεθοδολογίας ανάδυσης δεδομένων από τον Κρυμμένο Ιστό, οπότε όσες προσεγγίσεις περιγράφηκαν εκεί εφαρμόζονται και εδώ.

Τέλος στο βήμα της συγχώνευσης των αποτελεσμάτων, από τις εγγραφές που επιστρέφονται από τις κατά τόπους βάσεις αναγνωρίζονται όσες αντιστοιχούν στις ίδιες οντότητες, συνδυάζονται και παρουσιάζονται στους χρήστες συνήθως με κάποια κατάταξη. Με το βήμα αυτό είναι δυνατός ο συνδυασμός πληροφοριών από πολλαπλές πηγές για την ίδια οντότητα, κάτι που καθιστά δυνατή την παρουσίαση περισσότερων πληροφοριών στα αποτελέσματα. Για παράδειγμα, αν στην εγγραφή για ένα βιβλίο που προέρχεται από μια πηγή λείπει η τιμή για το πεδίο του συγγραφέα, και αυτή η τιμή βρίσκεται στην εγγραφή για το ίδιο βιβλίο από μια άλλη πηγή, τότε η ενοποιημένη εγγραφή θα είναι πλήρης. Επίσης είναι δυνατή η επισήμανση των διαφορών στις εγγραφές για την ίδια οντότητα που λαμβάνονται από διαφορετικές πηγές. Για παράδειγμα σε εφαρμογές σύγκρισης προϊόντων, οι χρήστες μπορεί να ενδιαφέρονται για τη σύγκριση των τιμών για το ίδιο προϊόν. Τέλος καθίσταται δυνατή η αφαίρεση των ίδιων εγγραφών που λαμβάνονται από διαφορετικές πηγές.

3.4.2.1 Πλεονεκτήματα και μειονεκτήματα

Με την άμεση προώθηση των ερωτημάτων των χρηστών στις μεμονωμένες βάσεις για την αναζήτηση αποτελεσμάτων, η μέθοδος της μετα-αναζήτησης επιφέρει πολλαπλά πλεονεκτήματα για τους χρήστες έναντι του μοντέλου της συλλογής δεδομένων:

- Το πρόβλημα της αδυναμίας προσκομιδής όλων των αποτελεσμάτων από πηγές του Κρυμμένου Ιστού δεν υπάρχει στη μέθοδο της μετα-αναζήτησης αφού αυτή έχει πρόσβαση σε όλα τα δεδομένα των πηγών της.
- Δεν τίθεται ζήτημα για το πόσο επίκαιρα είναι τα αποτελέσματα, αφού δεν υπάρχει κάποια φάση προσκομιδής δεδομένων. Αυτό σημαίνει ότι οι χρήστες θα έχουν πάντα πρόσβαση στα πιο πρόσφατα δεδομένα.
- Αυτή η προσέγγιση εκμεταλλεύεται στο έπακρο την υποδομή των βάσεων δεδομένων που υποστηρίζει, κάτι που σημαίνει ότι έχει χαμηλότερο κόστος για να υλοποιηθεί σε σύγκριση με τη μέθοδο της συλλογής δεδομένων.
- Η προσέγγιση αυτή παρέχει έναν πιο φυσικό τρόπο προσπέλασης των δεδομένων από τις κατά τόπους βάσεις, ακριβώς δηλαδή όπως κάθε μια από αυτές είχε σχεδιαστεί να χρησιμοποιείται.

Από την άλλη, η προσέγγιση αυτή παρουσιάζει και κάποια μειονεκτήματα:

- Αφού πραγματοποιούνται τόσα πολλά βήματα κατά την υποβολή ενός ερωτήματος, οι χρήστες μπορεί να μη λάβουν τόσο γρήγορα τα αποτελέσματα όσο στις υπόλοιπες μηχανές αναζήτησης.
- Αρκετές πηγές δεδομένων μπορεί να μην ανήκουν σε μια θεματική κατηγορία μόνο, ή μπορεί η κατηγορία αυτή να μην είναι ξεκάθαρη.
- Με αυτήν την προσέγγιση είναι αδύνατον να συνδεθούν κάποια δεδομένα με δεδομένα άλλων τύπων, κάτι που περιορίζει τη χρήση τους σε κάποιες εφαρμογές.
- Όταν αλλάζουν μεμονωμένα κάποιες βάσεις το σύστημα θα πρέπει να προβαίνει στις κατάλληλες ενημερώσεις για να μπορεί να λειτουργεί σωστά.

3.5 Το πρόβλημα της αναγνώρισης διεπαφών αναζήτησης

Η αναγνώριση των διεπαφών αναζήτησης είναι μια διαδικασία κοινή για κάθε σύστημα ανάδυσης ή μέτα-αναζήτησης, ενώ κατέστη σαφές από την παράγραφο 3.4 ότι αποτελεί αναγκαία προϋπόθεση για όλες τις υπόλοιπες διαδικασίες τέτοιου είδους συστημάτων. Δεδομένης της αχανούς, δυναμικής και αποκεντρωμένης φύσης του Παγκόσμιου Ιστού, η αυτοματοποίηση της αναγνώρισης των διεπαφών αναζήτησης είναι επιτακτική.

Το πρόβλημα της αναγνώρισης των διεπαφών αναζήτησης είναι το δεύτερο από τα δυο βήματα της διαδικασίας ανακάλυψης διεπαφών αναζήτησης στον Παγκόσμιο Ιστό (search interface discovery). Το πρώτο βήμα της διαδικασίας αυτής είναι ο εντοπισμός των διεπαφών στον Επιφανειακό Ιστό που εξαιτίας της μεγάλης διασποράς τους, παρουσιάζει αρκετές δυσκολίες. Η αναγνώριση διεπαφών αναζήτησης περιλαμβάνει το φιλτράρισμα όσων διεπαφών δεν προορίζονται για αναζήτηση, όπως τις διεπαφές εγγραφής ή σύνδεσης σε μια υπηρεσία ή τις διεπαφές ανάρτησης σχολίων σε ένα άρθρο. Άλλη μια πρόκληση για τη διαδικασία της αναγνώρισης των διεπαφών είναι η ίδια η μορφολογία των διεπαφών αυτών. Η συντριπτική πλειονότητα των διεπαφών στον Παγκόσμιο Ιστό είναι σε μορφή HTML κάτι που σημαίνει ότι (τουλάχιστον μέχρι την έκδοση HTML5) δεν φέρουν σημασιολογικές πληροφορίες που θα μπορούσαν να χρησιμοποιηθούν από ειδικά προγράμματα επεξεργασίας φορμών. Το κύριο μέλημα των προγραμματιστών κατά τη δημιουργία των διεπαφών είναι το πώς αυτές θα παρουσιάζονται στον τελικό χρήστη. Κατά συνέπεια τα αυτόματα εργαλεία εντοπισμού τους θα πρέπει να λαμβάνουν υπόψη μόνο τα συντακτικά στοιχεία τους.

Σύμφωνα με την εργασία [89], το πρόβλημα της ανακάλυψης διεπαφών αναζήτησης αποτελείται από τρεις διαστάσεις: α) τη διάσταση της θεματικής κατηγορίας και της λειτουργικότητας των διεπαφών, β) τη διάσταση της δομής τους και γ) τον τρόπο αναγνώρισής τους. Σύμφωνα με την πρώτη διάσταση στόχος των αυτόματων προγραμμάτων ανακάλυψης διεπαφών μπορεί να είναι η εύρεση της θεματικής κατηγορίας στην οποία ανήκουν (π.χ. βιβλία, αυτοκίνητα) ή η λειτουργικότητά

τους (διεπαφές αναζήτησης ή όχι). Σύμφωνα με τη δεύτερη διάσταση, τα προγράμματα ανακάλυψης διεπαφών μπορεί να επικεντρώνονται στις σύνθετες δομημένες διεπαφές ή τις απλές που βασίζονται σε λέξεις-κλειδιά. Στην πρώτη περίπτωση οι διεπαφές αποτελούνται από πολλαπλά πεδία που σχηματίζουν ένα υποτυπώδες σχήμα που παρέχει ενδείξεις για τη δομή της υποκείμενης βάσης δεδομένων. Στη δεύτερη περίπτωση εντάσσονται οι διεπαφές που αποτελούνται μόνο από ένα πεδίο το οποίο μάλιστα είναι κειμενικό και κατά συνέπεια παρέχει ελάχιστες έως καθόλου πληροφορίες σχετικά με την υποκείμενη βάση. Τέλος η τρίτη διάσταση διακρίνει τις προσεγγίσεις ανακάλυψης διεπαφών αναζήτησης, στις «προ-ερωτηματικές τεχνικές» (pre-query techniques), δηλαδή αυτές που επεξεργάζονται τη διεπαφή και μόνο αυτή για να λάβουν τις αποφάσεις τους σχετικά με την αναγνώριση και τις «μετ-ερωτηματικές τεχνικές» (post-query techniques), δηλαδή αυτές που υποβάλλουν ερωτήματα στη διεπαφή για να ενισχύσουν τις αποφάσεις τους με τα αποτελέσματα που αυτές επιστρέφουν. Οι πρώτες προσεγγίσεις είναι μεν απλούστερες στην υλοποίησή τους, αλλά έχουν λιγότερες πληροφορίες για να τεκμηριώσουν τις αποφάσεις τους, ενώ οι δεύτερες πρέπει να αντιμετωπίσουν πολύπλοκα ζητήματα, όπως την αυτόματη επιλογή τιμών για τα πεδία των διεπαφών κατά την υποβολή ερωτημάτων.

3.5.1 Κατηγοριοποίηση διεπαφών κατά λειτουργία

Σύμφωνα με την πρώτη διάσταση του προβλήματος της ανακάλυψης διεπαφών, υπάρχουν δυο είδη αναγνώρισης διεπαφών: α) η αναγνώριση κατά θεματική κατηγορία (identification by domain) και β) η αναγνώριση κατά λειτουργία (identification by function). Το πρώτο είδος εντάσσει τις διεπαφές σε θεματικές κατηγορίες, ανάλογα με το περιεχόμενο των πηγών δεδομένων με τις οποίες συσχετίζονται. Το δεύτερο είδος, διαχωρίζει τις διεπαφές ανάλογα με τη λειτουργία τους και πιο συγκεκριμένα σε αυτές που πραγματοποιούν αναζήτηση σε κάποια πηγή δεδομένων και σε οποιεσδήποτε άλλες. Τόσο το πρόβλημα της αναγνώρισης κατά λειτουργία όσο και το πρόβλημα τις αναγνώρισης κατά θεματική κατηγορία, μπορούν να αναχθούν σε προβλήματα κατηγοριοποίησης (classification problem).

Στο πρόβλημα της κατηγοριοποίησης διεπαφών κατά λειτουργία, στόχος είναι να οριστεί μια συνάρτηση $L: X \rightarrow Y$ που καλείται κατηγοριοποιητής, όπου

- $X = x_1, x_2, \dots x_n,$ είναι ένα σύνολο διεπαφών αναπαριστώμενων ως διανυσμάτων εισόδου, με $x_i = a_1, a_2, \dots a_k$ όπου a_j μια ιδιότητα.
- Y είναι το σύνολο των κατηγοριών, στην προκειμένη περίπτωση αν μια διεπαφή είναι αναζήτησης ή όχι.

Ένα από τα κύρια ζητήματα στα προβλήματα κατηγοριοποίησης είναι η επιλογή των κατάλληλων ιδιοτήτων πάνω στις οποίες θα βασιστεί η διαδικασία της μάθησης και η επιλογή του αλγόριθμου με τον οποίο θα γίνει η επαγωγή του κατηγοριοποιητή. Στις περισσότερες προσεγγίσεις στη

βιβλιογραφία όπως παρουσιάζονται στην παράγραφο 3.5.2, το πρόβλημα της κατηγοριοποίησης διεπαφών αντιμετωπίζεται με χειροκίνητα κατασκευασμένους ευρετικούς κανόνες, ενώ στις λιγοστές εργασίες που γίνεται χρήση κάποιου αλγόριθμου μηχανικής μάθησης, είτε χρησιμοποιούνται κειμενικές ιδιότητες, γεγονός που καθιστά το παραγόμενο κατηγοριοποιητή εφαρμόσιμο μόνο σε ιστοτόπους μιας συγκεκριμένης γλώσσας, είτε ο κατηγοριοποιητής δεν είναι εύκολα ερμηνεύσιμος από τον ανθρώπινο παράγοντα.

Στις επόμενες παραγράφους εξετάζεται α) η χρήση ιδιοτήτων ανεξάρτητων από τη γλώσσα συγγραφής της διεπαφής (ή της σελίδας που αυτή περιέχεται) για τη διαδικασία της κατασκευής του κατηγοριοποιητή και β) η χρήση του αλγόριθμου επαγωγής κανόνων που παρουσιάστηκε στο Κεφάλαιο 2 για τη δημιουργία μιας λίστας κανόνων για κατηγοριοποίηση.

Η αρχική επιλογή των ιδιοτήτων περιλαμβάνει το πλήθος εμφάνισης όλων των πιθανών στοιχείων εισόδου σε μια διεπαφή, όπως ορίζονται από την προδιαγραφή της HTML [90]. Πιο συγκεκριμένα οι ιδιότητες αυτές είναι το πλήθος εμφάνισης των δεκαεφτά στοιχείων εισόδου διεπαφών, τύπου κειμένου, κωδικού, κουμπιών κατάστασης (radio buttons), κουμπιών υποβολής, κουμπιών επαναφοράς, αρχείων, κρυμμένων πεδίων, εικόνων, απλών κουμπιών, κουμπιών με εικόνες, επιλογών, περιοχών κειμένου, επιλογών, ετικετών, συνόλου πεδίων και επικεφαλίδων για αυτά. Στην παράγραφο 3.6 αναλύεται το σύνολο διεπαφών Yahoo L11 με στόχο την ανακάλυψη περεταίρω πληροφοριών για τη συμπεριφορά των προαναφερθέντων ιδιοτήτων αλλά και την ενδεχόμενη εύρεση νέων ιδιοτήτων κατάλληλων για τη διαδικασία της μάθησης.

Για την εξαγωγή συμπερασμάτων σαν αλγόριθμος κατηγοριοποίησης χρησιμοποιείται ο αλγόριθμος επαγωγής κανόνων που παρουσιάστηκε στο κεφάλαιο Κεφάλαιο 2. Η επαγωγή κανόνων είναι γενικότερα ιδιαίτερα χρήσιμη στην ανάλυση δεδομένων και την ανακάλυψη γνώσης από δεδομένα, όπου η αναπαράσταση γνώσης χρειάζεται να γίνει με μέσα τα οποία είναι εύκολα ερμηνεύσιμα από τους ανθρώπους. Κατά συνέπεια μπορεί να φανεί χρήσιμη και στο πρόβλημα της κατηγοριοποίησης διεπαφών κατά λειτουργία.

3.5.2 Σχετική βιβλιογραφία

Σε μια από τις πιο παλιές εργασίες στο θέμα τις αναγνώρισης διεπαφών αναζήτησης στον Ιστό [91], οι συγγραφείς περιγράφουν έναν πράκτορα για αυτοματοποιημένη πραγματοποίηση ηλεκτρονικών αγορών με το όνομα ShopBot. Ο συγκεκριμένος πράκτορας μαθαίνει πώς να εξάγει πληροφορίες από ηλεκτρονικά καταστήματα χρησιμοποιώντας ελάχιστη γνώση πάνω στις θεματικές περιοχές των προϊόντων. Το υποσύστημα προσκομιδής δεδομένων του ShopBot, αρχικά βρίσκει τις διεπαφές που υπάρχουν σε έναν ιστότοπο, στη συνέχεια αναγνωρίζει ποιές από αυτές προορίζονται για αναζήτηση και τελικά επιλέγει τις καταλληλότερες για να τις αναλύσει «μετ-ερωτηματικά» για να βρει προϊόντα για σύγκριση. Η αναγνώριση των διεπαφών γίνεται με τη χρήση ενός φίλτρου που περιέχει έναν ευρετικό κανόνα ο οποίος διατυπώνεται ως εξής: «Οι διεπαφές αναζήτησης δεν

περιέχουν μερικές λέξεις που περιέχονται συχνά σε διεπαφές μη αναζήτησης όπως τηλέφωνο, e-mail, διεύθυνση» κλπ. Ωστόσο παρόλο που η υπόθεσή αυτή είναι λογική, δεν αξιολογείται με πειραματική μελέτη. Ακόμα και αν υπήρχε κάποιο λεξικό που περιέχει τέτοιες λέξεις, θα προορίζονταν για μια συγκεκριμένη γλώσσα και κατά συνέπεια θα χρειάζονταν διαφορετικά λεξικά, για διαφορετικές γλώσσες.

Στην εργασία [92], οι συγγραφείς στην προσέγγισή τους με το όνομα HiWE, ακολουθούν τη μεθοδολογία της ανάδυσης. Η προσέγγισή τους περιλαμβάνει έναν εξειδικευμένο προσκομιστή δεδομένων από τον Κρυμμένο Ιστό, που μπορεί να εξάγει πληροφορίες από ιστοτόπους. Ο προσκομιστής συμπληρώνει όσες διεπαφές έχει αναγνωρίσει ως αναζήτησης, από ένα σύνολο προκαθορισμένων τιμών και αποθηκεύει τα αποτελέσματα σε ένα αποθετήριο. Ο προσκομιστής μπορεί να συμπληρώσει μόνο τα ακόλουθα πεδία: πεδία κειμένου, λίστες επιλογών, κουμπιά radio και checkboxes. Παρόμοια με την εργασία [91], το HiWE χρησιμοποιεί ένα φίλτρο που βασίζεται σε ευρετικούς κανόνες, για να αναγνωρίσει τη λειτουργικότητα των διεπαφών αλλά και τη θεματική περιοχή τους. Ο πρώτος κανόνας στοχεύει στην αγνόηση, των διεπαφών αναζήτησης γενικού σκοπού, όπως αυτές που περιέχουν μόνο ένα κειμενικό πεδίο: «οι διεπαφές ερωτημάτων περιέχουν περισσότερα πεδία από ένα προκαθορισμένο ακέραιο αριθμό». Ο δεύτερος κανόνας στοχεύει στην επεξεργασία μόνο των διεπαφών εκείνων που ανήκουν σε μια θεματική κατηγορία: «οι διεπαφές ερωτημάτων περιέχουν πεδία των οποίων οι ετικέτες έχουν ταυτιστεί με κάποιες από ένα σύνολο προκαθορισμένων ετικετών στη βάση γνώσης». Η βάση γνώσης που κατασκευάζεται χειροκίνητα, περιέχει σύνολα ετικετών που είναι πιθανό να αντιστοιχιστούν με τις ετικέτες των πεδίων των διεπαφών. Κατά συνέπεια το HiWE φιλτράρει όσες φόρμες δεν περιέχουν το ελάχιστο πλήθος πεδίων ή που δεν έχουν αντιστοιχίσει τις ετικέτες τους με αυτές της βάσης γνώσης. Οι συγγραφείς δεν παρείχαν κάποια αξιολόγηση πάνω στην επίδοση των παραπάνω κανόνων, αφού ο κύριος στόχος τους ήταν η εξαγωγή του περιεχομένου στον Κρυμμένο Ιστό.

Άλλη μια εργασία που βασίζεται σε ευρετικούς κανόνες για να αναγνωρίσει τις διεπαφές αναζήτησης είναι η [74]. Σε αυτή προτείνεται ένας προσκομιστής δεδομένων εξειδικευμένος σε μια θεματική κατηγορία, που βασίζεται σε προ-κατηγοριοποιημένα κείμενα και επιλεγμένες λέξεις-κλειδιά, με απώτερο στόχο την ανακάλυψη διεπαφών αναζήτησης με ένα μόνο πεδίο. Οι συγγραφείς της εργασίας ορίζουν τέσσερα υποσύνολα διεπαφών: α) όλες τις πιθανές διεπαφές, β) τις κειμενικές διεπαφές, γ) τις διεπαφές ερωτημάτων και δ) τις διεπαφές αναζήτησης στον Κρυμμένο Ιστό. Ο προσκομιστής τους αναγνωρίζει τις διεπαφές ερωτημάτων εφαρμόζοντας διάφορους ευρετικούς κανόνες για να απορρίψει τις υπόλοιπες. Οι συγγραφείς αναφέρουν ορισμένους από αυτούς τους κανόνες: απορρίπτουν διεπαφές με μικρά κειμενικά πεδία (6 χαρακτήρες ή λιγότερους) και διεπαφές που περιέχουν πεδία κωδικών. Ωστόσο δεν παρέχουν την πλήρη λίστα με τους κανόνες αυτούς. Ο κύριος στόχος των συγγραφέων είναι η ανακάλυψη των διεπαφών αναζήτησης στον Κρυμμένο Ιστό, κάτι που απαιτεί περεταίρω ανάλυση και διερεύνηση με υποβολή ερωτημάτων και για το λόγο αυτό,

η αξιολόγηση που παρέχουν αφορά μόνο τη διαδικασία αυτή. Αυτό σημαίνει, ότι η επίδοση των ευρετικών κανόνων δεν είναι ξεκάθαρη. Οι συγγραφείς παρείχαν δυο σύνολα πειραμάτων που περιλάμβαναν την προσκομιδή σελίδων από 14 κατηγορίες ανώτερου επιπέδου από τον κατάλογο της Google, και μια τυχαία προσκομιδή σελίδων από τον Παγκόσμιο Ιστό. Τα αποτελέσματά τους για το πρώτο σύνολο, δείχνουν ότι ο προσκομιστής μπόρεσε να ανακαλύψει 5000 διεπαφές αναζήτησης από ένα σύνολο 25.000, ενώ για το δεύτερο σύνολο σχεδόν 20.000 κειμενικές διεπαφές από τις οποίες οι 7000 ήταν διεπαφές ερωτημάτων και 5000 ήταν διεπαφές αναζήτησης. Αυτοί οι αριθμοί καταδεικνύουν ότι ο προσκομιστής πετυχαίνει μια σταθερή αναλογία ανακαλυφθέντων διεπαφών και στα δύο πειράματα.

Στην εργασία [93], οι συγγραφείς μελετούν το πρόβλημα της συσταδοποίησης μηχανών αναζήτησης ηλεκτρονικών καταστημάτων μέσω των διεπαφών αναζήτησής τους, έτσι ώστε να κατασκευάσουν ένα σύστημα που υποστηρίζει ενοποιημένη πρόσβαση σε πολλαπλά ηλεκτρονικά καταστήματα. Η συσταδοποίηση στοχεύει στην ομαδοποίηση διεπαφών ανάλογα με τη θεματική περιοχή στην οποία ανήκουν. Τέτοιες προσεγγίσεις χρησιμοποιούν μέτρα ομοιότητας διεπαφών και πραγματοποιούν πολλαπλές συγκρίσεις ώστε να τοποθετήσουν κάθε διεπαφή στην ομάδα με τις πιο όμοιες τις. Η προσέγγισή τους με το όνομα WISE, αρχικά φιλτράρει τις διεπαφές που δεν προορίζονται για αναζήτηση. Όμοια με τις προηγούμενες προσεγγίσεις, χρησιμοποιούν ένα σύνολο ευρετικών κανόνων: α) η διεπαφή πρέπει να έχει την ιδιότητα action, β) η διεπαφή πρέπει να έχει ένα κουμπί υποβολής που περιέχει μια από τις λέξεις "search", "find", "query", "quote", γ) η διεπαφή πρέπει να έχει τουλάχιστον ένα πεδίο κειμένου ή μια λίστα επιλογών και δ) οι διεπαφές που έχουν μόνο checkboxes δεν είναι διεπαφές αναζήτησης. Αφού εφαρμοστούν αυτοί οι κανόνες, οι εναπομείνασες διεπαφές θεωρούνται υποψήφιες διεπαφές αναζήτησης. Οι συγγραφείς εξέτασαν οχτώ μεγάλες υποκατηγορίες σχετικές με το ηλεκτρονικό εμπόριο από τον κατάλογο της Yahoo και εξήγαγαν σχεδόν 300 διεπαφές από ένα σύνολο 270.000 σελίδων. Αυτές οι διεπαφές έχουν υποθετικά περάσει τους παραπάνω ελέγχους, ωστόσο οι συγγραφείς δεν αναφέρουν, πόσες απέτυχαν και πόσες από τις διεπαφές αναζήτησης δεν αναγνωρίστηκαν.

Η εργασία [94] είναι (με κάθε επιφύλαξη) η μοναδική που περιέχει μια αναλυτική λίστα ευρετικών κανόνων, η οποία υπόκειται σε πειραματικούς ελέγχους για τη δοκιμή της αξιοπιστίας της. Η λίστα αυτή έχει δυο στόχους: α) τη χρήση της σαν ατομική προσέγγιση για την κατηγοριοποίηση διεπαφών ως προς τη λειτουργία τους και β) τη χρήση της σε συνδυασμό με άλλους κατηγοριοποιητές για τη βελτίωση της απόδοσής τους. Η λίστα αποτελείται από τέσσερις κανόνες που αποτελούν όλους τους πιθανούς συνδυασμούς των ακόλουθων τριών συνθηκών: α) η διεπαφή χρησιμοποιεί ως μέθοδο υποβολής τη GET, β) η διεπαφή δεν περιέχει ένα πεδίο κωδικού και γ) η διεπαφή περιέχει τη λέξη "search". Οι συγγραφείς δοκιμάζουν την προτεινόμενη λίστα κανόνων σε συνδυασμό με έναν έτοιμο κατηγοριοποιητή και αναφέρουν ότι η απώλεια σε ακρίβεια ήταν αρκετά

μικρή, κάτι που σημαίνει ότι η λίστα μπορεί να χρησιμοποιηθεί σαν φίλτρο για την τροφοδότηση με δεδομένα εκπαίδευσης σε κατηγοριοποιητές.

Μια παρόμοια προσέγγιση ακολουθείται στην εργασία [95], με μόνη τη διαφορά ότι οι συγγραφείς δημιούργησαν τη λίστα αποκλειστικά για να βελτιώσουν την απόδοση των αλγόριθμων κατηγοριοποίησης. Οι συγγραφείς παρατηρώντας τις διεπαφές που φιλτραρίστηκαν με τη λίστα αυτή από ένα σύνολο δεδομένων, όρισαν ένα σύνολο ιδιοτήτων για δύο σύνολα εκπαίδευσης που περιέχουν μόνο τις ιδιότητες που εμφανίζονται περισσότερο στις φόρμες που πέρασαν από τους κανόνες. Οι φόρμες που προκύπτουν, αφού πρώτα μετατραπούν σε διανύσματα, κατηγοριοποιούνται χειροκίνητα και παρέχονται σε μια σειρά αλγόριθμων κατηγοριοποίησης και πιο συγκεκριμένα τους Naïve Bayes, J48, και SVM. Οι αλγόριθμοι αυτοί είχαν χρησιμοποιηθεί και στην εργασία [96]. Οι συγγραφείς παρατηρούν ότι οι επιδόσεις των αλγόριθμων βελτιώθηκαν σημαντικά, με ένα ποσοστό της τάξης του 6% σε σύγκριση με τις επιδόσεις που αναφέρονται στην εργασία [96].

Στην εργασία [97], οι συγγραφείς περιγράφουν έναν κατηγοριοποιητή διεπαφών ο οποίος βασίζεται στον αλγόριθμο επαγωγής δέντρων κατηγοριοποίησης C4.5 για την κατηγοριοποίηση των διεπαφών σε δυο κατηγορίες: αυτές που προορίζονται για αναζήτηση και όλες τις υπόλοιπες. Οι ιδιότητες για την αναπαράσταση των φορμών, παρήχθησαν αυτόματα από ένα δεδομένο σύνολο εκπαίδευσης (κάτι που σημαίνει ότι θα παραχθούν διαφορετικές ιδιότητες για διαφορετικά σύνολα δεδομένων) και περιλαμβάνουν δυαδικές ιδιότητες ανάλογα με τον τύπο των πεδίων που υπάρχουν σε μια διεπαφή και το σύνολο των λέξεων που έχουν βρεθεί σε συγκεκριμένα σημεία του κώδικα HTML: α) οι παράμετροι όνομα των ετικετών των πεδίων και η τιμή τους β) η παράμετρος όνομα των ετικετών των διεπαφών, και γ) το σύνολο των λέξεων από την παράμετρο action των ετικετών των διεπαφών. Η κατασκευή του δέντρου βασίστηκε σε δυο διαφορετικά σύνολα εκπαίδευσης καθένα από τα οποία περιείχε σχεδόν 200 φόρμες, 150 από τις οποίες χρησιμοποιήθηκαν για την εκπαίδευση και 50 από τις οποίες για την αξιολόγηση της ακρίβειας του δέντρου. Το πρώτο σύνολο συντέθηκε από έναν μόνο ονοματοχώρο, ενώ το δεύτερο συλλέχθηκε από τυχαίες σελίδες του Ιστού. Δυο σημαντικά ευρήματα προέκυψαν από τις πειραματικές μελέτες τους: α) το δέντρο που προέκυψε από το πρώτο σύνολο έχει καλύτερη ακρίβεια από αυτήν του δεύτερου και β) το δέντρο που προέκυψε από το δεύτερο σύνολο αποδίδει εξίσου καλά και στα δύο σύνολα. Έτσι τονίζεται το γεγονός ότι υπάρχει μια ισορροπία ανάμεσα στην ακρίβεια και τη γενικότητα των δέντρων, άρα η επιλογή του συνόλου εκπαίδευσης θα πρέπει να γίνει ανάλογα με το τι από τα δυο είναι πιο επιθυμητό. Ωστόσο, επειδή οι ιδιότητες αφορούν μόνο τα κειμενικά χαρακτηριστικά των διεπαφών, η προσέγγιση μπορεί να θεωρηθεί ότι εξαρτάται από τη γλώσσα, κάτι που σημαίνει ότι θα χρειαστεί δημιουργία πολλαπλών μοντέλων, ανάλογα με τις γλώσσες που είναι επιθυμητό να υποστηριχτούν.

Στην εργασία [96], οι συγγραφείς περιγράφουν έναν εστιασμένο προσκομιστή για τον εντοπισμό διεπαφών που ανήκουν σε μια θεματική κατηγορία. Ο προσκομιστής αποτελείται από τρεις κατηγοριοποιητές, έναν κατηγοριοποιητή σελίδων, έναν γενικό κατηγοριοποιητή διεπαφών και έναν

κατηγοριοποιητή θεματικών κατηγοριών. Ο κατηγοριοποιητής σελίδων χρησιμοποιεί ιδιότητες των σελίδων για να επικεντρωθεί σε ένα συγκεκριμένο θέμα και χρησιμοποιεί έναν απλό κατηγοριοποιητή Bayes για κείμενα, έτσι ώστε να αφαιρέσει διεπαφές που υπάρχουν σε άσχετες σελίδες. Στη συνέχεια ο γενικός κατηγοριοποιητής διεπαφών, αφαιρεί τις διεπαφές που δεν προορίζονται για αναζήτηση και τέλος ο κατηγοριοποιητής θεματικών κατηγοριών αναγνωρίζει τις διεπαφές αναζήτησης που ανήκουν σε μια θεματική κατηγορία, χρησιμοποιώντας επίσης έναν απλό κατηγοριοποιητή Bayes. Αυτά τα επίπεδα αφαιρέσεων αποσυνθέτουν το χώρο ιδιοτήτων με τέτοιον τρόπο που να είναι δυνατή η χρήση πιο κατάλληλων κατηγοριοποιητών για κάθε σύνολο ιδιοτήτων και να μη χρειάζεται ένας μοναδικός κατηγοριοποιητής που σίγουρα θα είχε χειρότερα αποτελέσματα. Αυτή η ιδέα είναι κεντρική και για ένα σύνολο άλλων εργασιών από τους ίδιους συγγραφείς [98], [99]. Ο γενικός κατηγοριοποιητής χρησιμοποιεί τον αλγόριθμο επαγωγής δέντρων C4.5 όπως στην εργασία [97], ωστόσο το πλήθος των ιδιοτήτων εδώ είναι προκαθορισμένο. Πιο συγκεκριμένα οι συγγραφείς χρησιμοποίησαν 16 αριθμητικές και δυαδικές ιδιότητες των διεπαφών, που αποτελούνται από τα πλήθη των διαφόρων πεδίων εισόδου των διεπαφών μαζί με τη μέθοδο υποβολής (GET/POST) και την παρουσία της λέξης "search". Η φάση της εκπαίδευσης βασίστηκε σε ένα σύνολο που αποτελούνταν από 475 διεπαφές, 216 από τις οποίες προήλθαν από το αποθετήριο UIUC [100] και προορίζονταν για αναζήτηση, ενώ 259 από τις οποίες προήλθαν από χειροκίνητη συλλογή από τους συγγραφείς και δεν προορίζονταν για αναζήτηση. Τα πειράματα περιελάμβαναν μια σύγκριση του C4.5 με τρεις άλλες τεχνικές μηχανικής μάθησης: τον απλό κατηγοριοποιητή Bayes, το πολυεπίπεδο perceptron (multi-layer perceptron) και τις μηχανές διανυσμάτων υποστήριξης (support vector machines), η οποία έδειξε ότι ο C4.5 είχε τη χαμηλότερη αναλογία σφαλμάτων (9.05%).

3.6 Το σύνολο δεδομένων Yahoo L11

Το σύνολο δεδομένων Webscope L11 [101] της Yahoo περιέχει ένα δείγμα πολύπλοκων διεπαφών ερωτημάτων. Πολύπλοκες θεωρούνται οι διεπαφές που έχουν τρία ή περισσότερα πεδία εισόδου, όπως πεδία κειμένου, κουτιά ελέγχου, εικόνες, επιλογών και κουμπιά. Το δείγμα περιέχει 2.67 εκατομμύρια τέτοιες διεπαφές. Τα δεδομένα του δείγματος αυτού μπορούν να φανούν χρήσιμα για μελέτες που αφορούν την κατηγοριοποίηση των φορμών και την ανάδυση δεδομένων από τον Κρυμμένο Ιστό. Σύμφωνα με τη Yahoo [102] το δείγμα αυτό συγκαταλέγεται στα σύνολα μεγάλης κλίμακας που διαθέτει δημόσια για μελέτη. Σύμφωνα όμως με το κεφάλαιο 1 της διατριβής αυτής, το δείγμα αυτό δεν μπορεί να θεωρηθεί ως σύνολο Μεγάλων Δεδομένων, καθώς μπορεί να αναλυθεί με συμβατικές μεθόδους επεξεργασίας δεδομένων. Η χρήση του ωστόσο σε συνδυασμό με τεχνικές επεξεργασίας Μεγάλων Δεδομένων σε αυτή τη διατριβή οφείλεται α) στο γεγονός ότι είναι το μοναδικό επαρκώς μεγάλο, δημόσια διαθέσιμο δείγμα για τη μελέτη των διεπαφών αναζήτησης, β)

στην ευκολία της αναπαραγωγής των μετρήσεων και της επιβεβαίωσής τους από τρίτους και γ) στο χαμηλό κόστος που χρειάζεται για την επεξεργασία του.

3.6.1 Δομή του συνόλου δεδομένων Yahoo L11

Το δείγμα Yahoo L11 αποτελείται από 30 αρχεία απλού κειμένου (plaintext) που είναι αποθηκευμένα στην υπηρεσία Amazon S3 και είναι διαθέσιμα κατόπιν αίτησης στη Yahoo μέσα από το πρόγραμμα Webscope. Κάθε αρχείο είναι ασυμπίεστο και το μέγεθός του κυμαίνεται στα 4.4GB κάτι που σημαίνει ότι το συνολικό δείγμα ανέρχεται στα 133GB. Κάθε αρχείο αποτελείται από πολλαπλές πλειάδες που αντιστοιχούν σε πολύπλοκες φόρμες. Κάθε πλειάδα έχει την εξής μορφή:

```
<url, form_id, form_action, content>
```

όπου url η διεύθυνση της σελίδας στην οποία ανήκει η φόρμα, form_id το αναγνωριστικό της φόρμας, form_action η διεύθυνση στην οποία υποβάλλεται η φόρμα και content ο κώδικας HTML της σελίδας που περιέχει τη φόρμα.

3.6.2 Χαρακτηριστικά προς ανάλυση

Για τους σκοπούς της στατιστικής μελέτης αυτής, εξήχθησαν ένα σύνολο χαρακτηριστικών για κάθε μια πλειάδα που αναλύθηκαν από την προοπτική δύο επιπέδων λεπτομέρειας: α) το επίπεδο σελίδας και β) το επίπεδο φόρμας HTML. Στο πρώτο επίπεδο, ανήκουν τα ακόλουθα χαρακτηριστικά:

- Το url, από το οποίο μπορούν να εξαχθούν ο ονοματοχώρος (domain name) του ιστότοπου, η δημόσια κατάληξή του (public suffix) και το όνομα του χώρου υψηλότερου επιπέδου (top level domain). Η δημόσια κατάληξη είναι το επίπεδο κάτω από το οποίο μπορεί να δεσμεύσει κανείς έναν ιδιωτικό χώρο. Για παράδειγμα η δημόσια κατάληξη του "primeminister.gov.gr" είναι η "gov.gr", ενώ το όνομα του κορυφαίου χώρου για τη συγκεκριμένη διεύθυνση είναι το "gr". Μια ενημερωμένη λίστα των κορυφαίων καταλήξεων παρέχεται από το Ίδρυμα Mozilla [103].
- Η γλώσσα της σελίδας. Για την αναγνώριση της γλώσσας στην οποία είναι γραμμένη μια σελίδα, χρησιμοποιήθηκε ένας απλός κατηγοριοποιητής Bayes [104], που βασίζεται στις πιθανότητες ενός συνόλου ν-γραμμάτων (n-grams) που έχουν οριστεί για κάθε υποστηριζόμενη γλώσσα.
- Η έκδοση της HTML. Με την HTML5 [105] εισήχθησαν εξειδικευμένες ετικέτες που εμπλουτίζουν σημασιολογικά τις σελίδες στις οποίες περιέχονται. Οι σημασιολογικές ετικέτες αυτές, πέραν των παρουσιαστικών διαφορών, περιγράφουν καλύτερα το περιεχόμενό τους. Επιπρόσθετα, έχει προστεθεί λειτουργικότητα η οποία στις προηγούμενες εκδόσεις της παρέχονταν μέσω της γλώσσας Javascript. Ένα τέτοιο παράδειγμα είναι η επικύρωση μιας εισαχθείσας από το χρήστη τιμής σε μια HTML

φόρμα. Αυτές οι ετικέτες λοιπόν, προσφέρουν οδηγίες για το πώς μπορούν να ερμηνευτούν τα περιεχόμενά τους, τις οποίες μπορεί να εκμεταλλευτεί λογισμικό όπως τα προγράμματα προσκομιδής σελίδων ή οι περιηγητές Ιστού. Στόχος της μέτρησης του χαρακτηριστικού αυτού, αποτελεί η ανάδειξη της χρησιμότητας του δείγματος Yahoo L11 για την ανάπτυξη μεθόδων ανάδυσης ή ενοποίησης που θα εκμεταλλεύονται τα χαρακτηριστικά της HTML5.

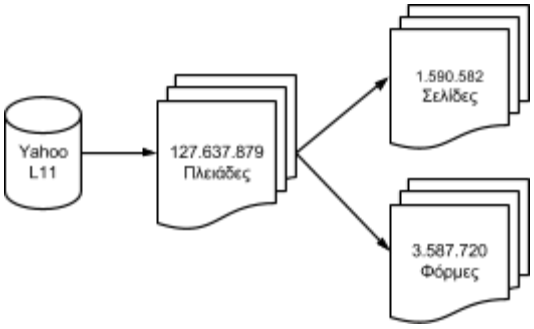
- Το πλήθος των φορμών ανά σελίδα. Το χαρακτηριστικό αυτό αποσκοπεί στην εύρεση της κατανομής των φορμών ανά σελίδα, έτσι ώστε να προσδιοριστεί ο φόρτος εργασίας που θα έχει ένα αυτοματοποιημένο πρόγραμμα προσκομιδής περιεχομένου κατά τη διαδικασία της εύρεσης διεπαφών αναζήτησης στον Επιφανειακό Ιστό.

Στο επίπεδο λεπτομέρειας που αφορά τις φόρμες HTML αναλύονται τα ακόλουθα χαρακτηριστικά:

- Η συχνότητα εμφάνισης των στοιχείων εισόδου. Στόχος της μέτρησης του χαρακτηριστικού αυτού, είναι να προσδιοριστεί πόσο συχνά εμφανίζονται οι διάφοροι τύποι εισόδου στις φόρμες που υπάρχουν στο δείγμα και να γίνει μια πρώτη εκτίμηση της σημασίας τους. Από τη συχνότητα εμφάνισης κάθε πεδίου μπορεί να γίνει επίσης μια υποτυπώδης χειροκίνητη διακριτοποίηση των τιμών των πεδίων, για να μπορούν να χρησιμοποιηθούν από ορισμένους αλγόριθμους μηχανικής μάθησης.
- Τα μέτρα κεντρικής τάσης και τα μέτρα διασποράς για κάθε στοιχείο φόρμας. Τα μέτρα κεντρικής τάσης (measures of central tendency) που μπορούν να χρησιμοποιηθούν σαν αντιπροσωπευτικά μέτρα περιγραφής του δείγματος. Με άλλα λόγια προσδιορίζουν την τυπικότητα του δείγματος. Πιο συγκεκριμένα χρησιμοποιήθηκαν τα τρία πιο συνηθισμένα μέτρα κεντρικής τάσης που είναι η μέση τιμή, η διάμεσος και η επικρατούσα τιμή. Τα μέτρα διασποράς (measures of variability) δίνουν μια εικόνα σχετικά με το πόσο συγκεντρωμένες είναι οι παρατηρήσεις σε ένα σύνολο δεδομένων. Το μοναδικό μέτρο που χρησιμοποιήθηκε είναι το ενδοτεταρτομοριακό εύρος που φαίνεται από τα θηκογράμματα των στοιχείων εισόδου.

3.6.3 Μεθοδολογία

Η Εικόνα 3.6 εμφανίζει τη ροή δεδομένων κατά την ανάλυση του δείγματος Yahoo L11. Το δείγμα αποτελείται από 127.637.879 πλειάδες, από την επεξεργασία των οποίων προέκυψαν δύο σύνολα δεδομένων: α) ένα σύνολο από 1.590.582 HTML σελίδες και β) ένα σύνολο από 3.587.720 φόρμες.



Εικόνα 3.6- Ροή δεδομένων της στατιστικής μελέτης του συνόλου δεδομένων Yahoo L11

Για τη μελέτη του δείγματος, αρχικά αναλύθηκαν οι πλειάδες στην αρχική μορφή τους και κρατήθηκαν το url και το περιεχόμενό τους, από τα οποία εξήχθησαν το όνομα χώρου του ιστοτόπου, η δημόσια κατάληξή του, το όνομα του κορυφαίου χώρου, η γλώσσα της σελίδας, η έκδοση της HTML, το πλήθος των φορμών ανά σελίδα και το πλήθος των στοιχείων εισόδου κάθε φόρμας. Η μορφή που επιλέχθηκε για τα δεδομένα αυτά είναι η μορφή τιμών διαχωριζόμενων από χαρακτήρες στηλοθετών (tab separated values). Η μορφή αυτή επιτρέπει την υποβολή ερωτημάτων τύπου SQL σε μια μηχανή όπως της Apache Hive [106].

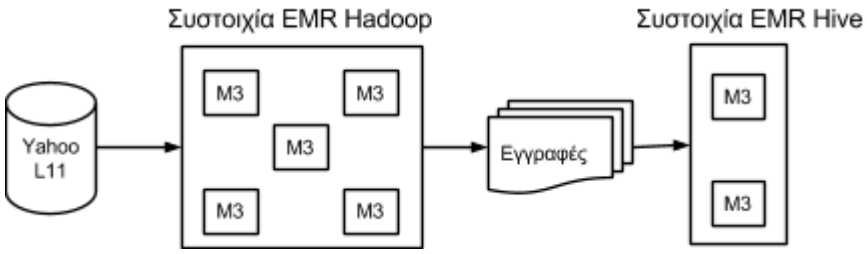
Η διαδικασία της μετατροπής των δεδομένων του Yahoo L11 από την αρχική τους μορφή σε τιμές διαχωριζόμενες από χαρακτήρες στηλοθετών, πραγματοποιήθηκε με τη γνωστή προσέγγιση επεξεργασίας Μεγάλων Δεδομένων Απεικόνισης/Μείωσης. Μια τέτοια επιλογή είναι επιβεβλημένη δεδομένου ότι υπάρχει η απαίτηση η διαδικασία να είναι κλιμακούμενη, όπως χρειάζεται άλλωστε στις περιπτώσεις όπου τα δεδομένα προς ανάλυση είναι όντως μεγάλης κλίμακας. Το έργο Απεικόνισης/Μείωσης για τη μετατροπή των δεδομένων περιγράφεται στον Αλγόριθμος 3.1.

Το έργο αυτό αποτελείται μόνο από μια εργασία απεικόνισης. Εφόσον δεν χρειάζεται να πραγματοποιηθεί κάποιο είδος συνάθροισης, η εργασία της απεικόνισης αρκεί για την πραγματοποίηση της μετατροπής και δεν χρειάζεται διαδικασία Μείωσης. Η συνάρτηση της απεικόνισης καλείται για κάθε γραμμή ενός δεδομένου αρχείου του δείγματος Yahoo L11. Η συνάρτηση αρχικά ελέγχει να βρει τους ειδικούς χαρακτήρες που διαμορφώνουν μια πλειάδα για να αποφασίσει το είδος της. Αν βρεθεί κενή πλειάδα ή πλειάδα με κενό περιεχόμενο τότε εκπέμπεται μια εγγραφή χωρίς τιμή, μόνο και μόνο για να καταμετρηθούν τα δυο είδη πλειάδων αυτά. Αν η γραμμή περιέχει τον ειδικό χαρακτήρα έναρξης μιας νέας πλειάδας με περιεχόμενο, τότε κάθε επόμενη γραμμή θα αποθηκεύεται σε μια τοπική μεταβλητή που θα συναθροίζει το περιεχόμενό της μέχρι να βρεθεί ο ειδικός χαρακτήρας τερματισμού. Αφού αποκτηθεί μια πλήρης πλειάδα με περιεχόμενο, εξάγονται τα χαρακτηριστικά που αφορούν τη σελίδα και τις φόρμες τις και εκπέμπονται στην έξοδο. Με το πέρας του συγκεκριμένου έργου θα έχει αποθηκευθεί στην έξοδο το σύνολο δεδομένων σε μορφή τιμών διαχωριζόμενων από στηλοθέτες. Τα δεδομένα αυτά προορίζονται για περαιτέρω ανάλυση και εξαγωγή στατιστικών.

Η εκτέλεση του παραπάνω έργου πραγματοποιήθηκε σε μια συστοιχία εικονικών μηχανών της υπηρεσίας EMR της Amazon, αποτελούμενη από 5 στιγμιότυπα τύπου m3.xlarge που παρείχαν 4 εικονικούς επεξεργαστές, 15GB μνήμης και 2x40GB αποθηκευτικού χώρου. Η διαδικασία της μετατροπής διήρκεσε 9 ώρες και κόστισε περίπου 3 ευρώ. Η αντίστοιχη εκτέλεση της διαδικασίας της μετατροπής σε μια μόνο μηχανή με κεντρική μέθοδο επεξεργασίας διήρκεσε 40 ώρες. Από τα δεδομένα αυτά είναι προφανές το πλεονέκτημα της χρήσης παράλληλης επεξεργασίας ακόμα και για αυτό το δείγμα. Μια περίληψη της διαδικασίας εκτέλεσης φαίνεται στην Εικόνα 3.7.

Αλγόριθμος 3.1 - Διαδικασία Απεικόνισης/Μείωσης για τη μετατροπή των ιστοσελίδων σε μορφή τιμών διαχωριζόμνων από χαρακτήρες στηλοθετών

```
Είσοδος: γραμμή
Έξοδος: <τύπος, εγγραφή>
1.τρέχουσα_πλειάδα = null
2.Συνάρτηση Απεικόνιση(κλειδί, γραμμή)
3.    Αν η γραμμή περιέχει κενή πλειάδα
4.       εκπομπή (τύπος_κενής, NULL)
5.    Αλλιώς Αν η γραμμή περιέχει πλειάδα με κενό περιεχόμενο
6.       εκπομπή (τύπος_κενό_περιεχόμενο, NULL)
7.    Αλλιώς αν δεν βρέθηκε η ένδειξη τερματισμού πλειάδας με περιεχόμενο
8.       τρέχουσα_πλειάδα += γραμμή
9.    Αλλιώς
10.          εγγραφή_σελίδας = εξαγωγή_χαρακτηριστικών(τρέχουσα_πλειάδα)
11.          Λίστα<εγγραφή_φόρμας> =
   εξαγωγή_χαρακτηριστικών(τρέχουσα_σελίδα)
12.          έκπεμπε(τύπος_σελίδας, εγγραφή_σελίδας)
13.          Για κάθε εγγραφή φόρμας στη Λίστα<εγγραφή_φόρμας)
14.        εκπομπή(τύπος_φόρμας, εγγραφή_φόρμας)
```



Εικόνα 3.7 - Η ροή εκτέλεσης της μετατροπής των δεδομένων του Yahoo L11

3.6.4 Αποτελέσματα στο επίπεδο σελίδας

3.6.4.1 Πλειάδες

Η πρώτη ανάλυση σχετικά με τα αποτελέσματα των μετρήσεων των προαναφερθέντων χαρακτηριστικών αφορά τις ίδιες τις πλειάδες. Στο δείγμα παρατηρήθηκαν ορισμένες αστοχίες από τους δημιουργούς του. Αρχικά εντοπίστηκαν αρκετές πλειάδες που ήταν κενές, δηλαδή δεν περιείχαν

τιμές για τα αντίστοιχα πεδία τους. Στη συνέχεια εντοπίστηκαν πολλαπλές πλειάδες που δεν διέθεταν το περιεχόμενο της σελίδας στο οποίο περιέχονται φόρμες. Οι δυο προαναφερθέντες τύποι πλειάδων δεν επηρέασαν το πραγματικό μέγεθος του δείγματος, αφού αυτό ουσιαστικά διαμορφώθηκε από το τελευταίο είδος πλειάδων. Το είδος αυτό αφορά τις πλειάδες που είχαν περιεχόμενο. Αρκετές από αυτές ωστόσο εμφανίζονταν πολλαπλές φορές κάτι που σημαίνει ότι στο δείγμα υπήρχαν διπλότυπα. Αυτό οφείλεται στο γεγονός ότι ορισμένες σελίδες περιείχαν πολλαπλές πολύπλοκες διεπαφές και επειδή η δομή της πλειάδας μπορεί να περιέχει το όνομα μόνο μιας διεπαφής, ο μοναδικός τρόπος αναφοράς σε όλες τις διεπαφές περιλάμβανε την ταυτόχρονη επανάληψη και του περιεχομένου. Ο Πίνακας 3.1 συνοψίζει το πλήθος των διαφόρων τύπων πλειάδων που βρέθηκαν στο δείγμα.

Πίνακας 3.1 - Σύνοψη των πλειάδων που βρέθηκαν στο σύνολο δεδομένων Yahoo L11

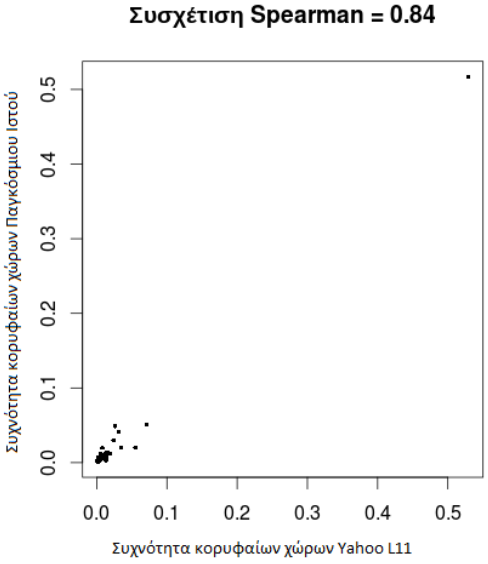
	Σύνολο	**Κενές**	**Χωρίς Περιεχόμενο**	**Με Περιεχόμενο**	**Μοναδικές**
Απόλυτη Συχνότητα	127.637.879	78.963.332	47.083.965	2.675.822	1.590.582
Ποσοστό	100%	61.86%	36.88%	2.09%	1.24%

3.6.4.2 Κορυφαίοι ονοματοχώροι

Το πρώτο χαρακτηριστικό που μετρήθηκε στο επίπεδο λεπτομέρειας σελίδων αφορά τα τους χώρους κορυφαίου επιπέδου. Ένα από τα κύρια ερωτήματα σχετικά με το χαρακτηριστικό αυτό είναι ποιοι χώροι υπάρχουν και ποια είναι η αναλογία τους στο συνολικό δείγμα. Για την απάντηση του ερωτήματος αυτού, οι δημόσιες καταλήξεις συγχωνεύτηκαν στο αντίστοιχο κορυφαίο επίπεδο (π.χ. η .gov.gr στο .gr). Στον Πίνακας 3.2 εμφανίζονται τα αποτελέσματα με σχετική συχνότητα πάνω από 1%. Για το δείγμα L11 υπάρχουν 15 χώροι κορυφαίου επιπέδου πάνω από το όριο του 1%. Από τον πίνακα είναι πρόδηλη η υπεροχή του χώρου .com που αντιστοιχεί στο μισό των εγγράφων του δείγματος. Αυτό οφείλεται στο γεγονός ότι ο ονοματοχώρος .com περιέχει ιστοτόπους από όλον τον κόσμο. Το υπόλοιπο μισό του δείγματος κατανέμεται στους υπόλοιπους ονοματοχώρους. Χρησιμοποιώντας αυτά τα ευρήματα και συγκρίνοντάς τα με την κατανομή των κορυφαίων ονοματοχώρων όπως παρέχονται από την επισκόπηση για τη χρήση των κορυφαίων ονοματοχώρων που βρίσκεται στο [107], είναι δυνατόν να υπολογιστεί μια εκτίμηση και το πόσο αντιπροσωπευτικό του Παγκόσμιου Ιστού είναι το δείγμα Yahoo L11 και αν υπάρχει κάποια προτίμηση σε κάποιον από τους χώρους κορυφαίου επιπέδου. Η εκτίμηση της αντιπροσώπευσης αυτής, ουσιαστικά αντιστοιχεί στην εκτίμηση της μονοτονικής σχέσης μεταξύ των ταξινομημένων σχετικών συχνοτήτων κάθε χώρου κορυφαίου επιπέδου, όπως δίνεται από το δείγμα Yahoo L11 και από την παραπάνω επισκόπηση. Για τον υπολογισμό της συσχέτισης αυτής μπορεί να χρησιμοποιηθεί ο συντελεστής συσχέτισης Spearman ο οποίος για ένα δείγμα μεγέθους n δυο μεταβλητών Xi,Yi ορίζεται ως:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \text{όπου } d_i = x_i - y_i \quad (3.1)$$

Ο συντελεστής συσχέτισης Spearman για τις δυο αυτές μεταβλητές ανήλθε στην τιμή 0.84 για τις πρώτες 50 τιμές, δείχνοντας θετική συσχέτιση μεταξύ των δυο μεταβλητών κάτι που αποτελεί μια καλή ένδειξη ότι το δείγμα Yahoo L11 είναι σχετικά αντιπροσωπευτικό δείγμα του Παγκόσμιου Ιστού.



Εικόνα 3.8 - Συσχέτιση Spearman για τις συχνότητες κορυφαίων χώρων του Yahoo L11 και του πραγματικού Ιστού

Πίνακας 3.2 - Πίνακας συχνοτήτων των κορυφαίων ονοματοχώρων του Yahoo L11

TLD	Απόλυτη Συχνότητα Yahoo L11	Σχετική Συχνότητα Yahoo L11
com	842640	0.5290
net	114018	0.0710
jp	88485	0.0550
uk	56365	0.0350
org	49573	0.0310
ru	41548	0.0260
de	38241	0.0240
it	31672	0.0190
pl	23876	0.0150
cn	23167	0.0140
tw	21899	0.0137
kr	21342	0.0134
es	20857	0.0131
nl	19295	0.0121
fr	17877	0.0112
Υπόλοιπα	179727	0.1129

3.6.4.3 Ονοματοχώροι

Στον Πίνακας 3.3 φαίνονται οι 10 πιο συχνοί ονοματοχώροι με κατάληξη com και στον Πίνακας 3.4, οι 10 πιο συχνοί ονοματοχώροι με οποιαδήποτε άλλη κατάληξη. Με το 6.9% των συνολικών σελίδων, ο ονοματοχώρος yahoo.com είναι ο πιο συχνά εμφανιζόμενος στο δείγμα Yahoo L11. Άλλοι ονοματοχώροι σε αυτή την κατηγορία περιλαμβάνουν ηλεκτρονικά καταστήματα, υπηρεσίες ιστολογίων, ταξιδιωτικούς και ειδησεογραφικούς ιστότοπους. Στην κατηγορία με τους ονοματοχώρους με οποιαδήποτε κατάληξη εκτός της .com (Πίνακας 3.4), φαίνεται ότι έξι ονοματοχώροι ανήκουν στις αμέσως πιο συχνές καταλήξεις πέραν της com, όπως φαίνονται στον Πίνακας 3.3. Πρώτος ονοματοχώρος σε αυτήν την κατηγορία είναι το ηλεκτρονικό κατάστημα ecplaza.net ενώ εδώ συμπεριλαμβάνονται ειδησεογραφικοί ιστότοποι και portals. Μια γενική παρατήρηση που εξάγεται από τα αποτελέσματα αυτά είναι ότι το δείγμα δεν στοχεύει κάποιες

συγκεκριμένες θεματικές κατηγορίες του Κρυμμένου Ιστού, ενώ φαίνεται ότι επιλέχθηκαν απλώς ιστότοποι με φόρμες που καλύπτουν το κριτήριο της πολυπλοκότητας όπως ορίστηκε προηγουμένως. Επίσης παρατηρείται η παντελής απουσία άλλων γνωστών ιστοτόπων στον ονοματοχώρο com (π.χ. amazon, youtube κ.λ.π) που διαθέτουν διεπαφές αναζήτησης.

Πίνακας 3.3 - Οι δέκα συχνότεροι ονοματοχώροι από τον κορυφαίο ονοματοχώρο com

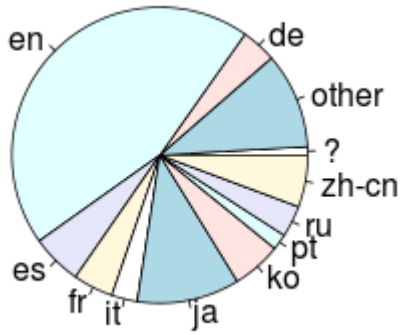
Σειρά	Ονοματοχώρος	Απόλυτη Συχνότητα	Σχετική Συχνότητα
1	yahoo.com	111227	0.0699
2	taobao.com	33320	0.0209
3	xanga.com	19988	0.0125
4	pantip.com	6835	0.0042
5	petfinder.com	6557	0.0041
6	globo.com	6091	0.0038
7	made-in-china.com	4739	0.0029
8	mundorecetas.com	4114	0.0025
9	hoteltravel.com	3629	0.0022
10	real.com	3536	0.0022

Πίνακας 3.4 – Οι δέκα συχνότεροι ονοματοχώροι από τους υπόλοιπους κορυφαίους ονοματοχώρους

Σειρά	Ονοματοχώρος	Απόλυτη Συχνότητα	Σχετική Συχνότητα
1	ecplaza.net	17673	0.0111
2	topix.net	4414	0.0027
3	yahoo.co.jp	3689	0.0023
4	auction.co.kr	3521	0.0022
5	biglobe.ne.jp	3372	0.0021
6	townwork.net	2832	0.0017
7	gismeteo.ru	2552	0.0016
8	elmundo.es	2549	0.0016
9	guardian.co.uk	2506	0.0015
10	ekikara.jp	2466	0.0015

3.6.4.4 Γλώσσες

Στον Πίνακας 3.5 φαίνεται η κατανομή των 10 πιο συχνών γλωσσών που εμφανίζονται στο δείγμα Yahoo L11. Η πρώτη παρατήρηση που προκύπτει από αυτά είναι ότι υπάρχει σημαντική διαφοροποίηση από την αντίστοιχη λίστα των χώρων ανώτατου επιπέδου, κάτι που οφείλεται προφανώς στην πολυγλωσσία που διακρίνει τους ιστότοπους στον χώρο com. Η δεύτερη παρατήρηση που προκύπτει είναι ότι υπάρχει ποικιλία στις εμφανιζόμενες γλώσσες κάτι που σημαίνει ότι το δείγμα μπορεί να αποτελέσει σημείο έναρξης για μελέτη του Κρυμμένου Ιστού μιας συγκεκριμένης χώρας.



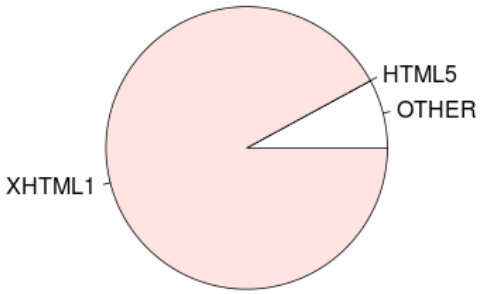
Εικόνα 3.9 - Κατανομή γλωσσών στο δείγμα Yahoo L11

Πίνακας 3.5 - Συχνότητα εμφάνισης γλωσσών

Γλώσσα	**Απόλυτη Συχνότητα**	**Σχετική Συχνότητα**
Αγγλικά (en)	709592	0.4461
Ιαπωνικά (jp)	178577	0.1122
Άλλες (other)	166732	0.1048
Κινέζικα (zh-cn)	90136	0.0566
Ισπανικά (es)	87741	0.0551
Κορεάτικα (ko)	85424	0.0537
Γαλλικά (fr)	69811	0.0438
Γερμανικά (de)	60965	0.0383
Ρώσικα (ru)	54766	0.0344
Ιταλικά (it)	44612	0.0280
Πορτογαλικά (pt)	27640	0.0173
Μη αναγνωρισμένη (?)	14586	0.0091

3.6.4.5 Κατανομή εκδόσεων HTML

Στην Εικόνα 3.10 φαίνεται ξεκάθαρα ότι υπάρχουν απειροελάχιστες σελίδες έκδοσης HTML5, ενώ η κυρίαρχη έκδοση είναι της XHTML1. Αυτό σημαίνει ότι το δείγμα Yahoo L11 δεν ενδείκνυται για μελέτες που εκμεταλλεύονται τα νέα χαρακτηριστικά της HTML5 για την ανάπτυξη μεθόδων πρόσβασης στα δεδομένα του Κρυμμένου Ιστού.

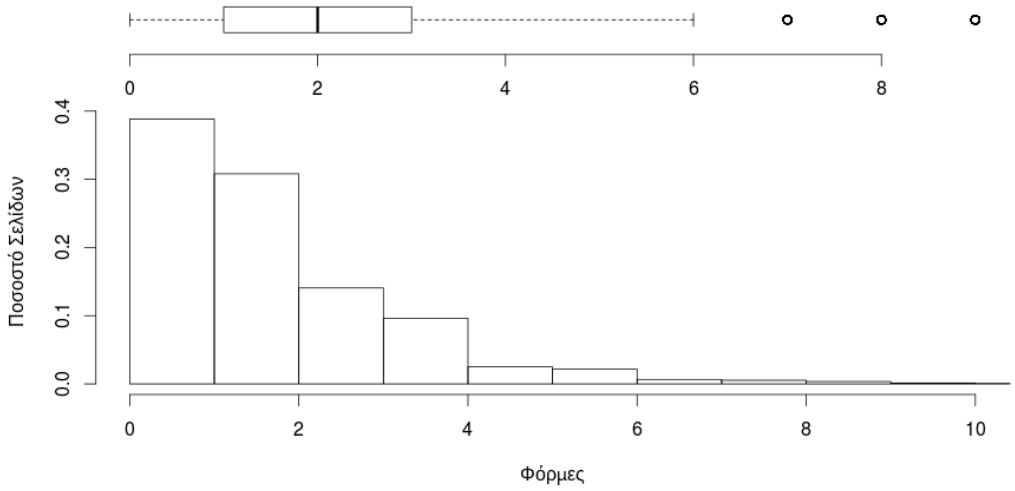


Εικόνα 3.10 - Κατανομή εκδόσεων HTML

3.6.4.6 Κατανομή διεπαφών ανά σελίδα

Στην Εικόνα 3.11 φαίνεται η κατανομή διεπαφών ανά σελίδα. Η μεγάλη πλειονότητα των σελίδων περιέχει μία ή δύο διεπαφές, ενώ σταδιακά η συχνότητα του πλήθους των διεπαφών ανά σελίδα μειώνεται μέχρι την τιμή 6, οπότε και ουσιαστικά μηδενίζεται. Στο δείγμα παρατηρήθηκαν

αρκετές ακραίες τιμές με μέγιστη την εμφάνιση 309 διεπαφών σε μια σελίδα. Ένα συνηθισμένο σενάριο που παρατηρήθηκε στο δείγμα, είναι μια σελίδα να περιλαμβάνει μια φόρμα σύνδεσης (login) στον ιστότοπο και μια φόρμα αναζήτησης, ενώ στις ακραίες περιπτώσεις, το ασυνήθιστο πλήθος φορμών οφείλεται είτε σε καθαρά προγραμματιστική επιλογή, είτε σε προγραμματιστική αστοχία.



Εικόνα 3.11 - Κατανομή διεπαφών ανά σελίδα

3.6.5 Αποτελέσματα στο επίπεδο διεπαφής

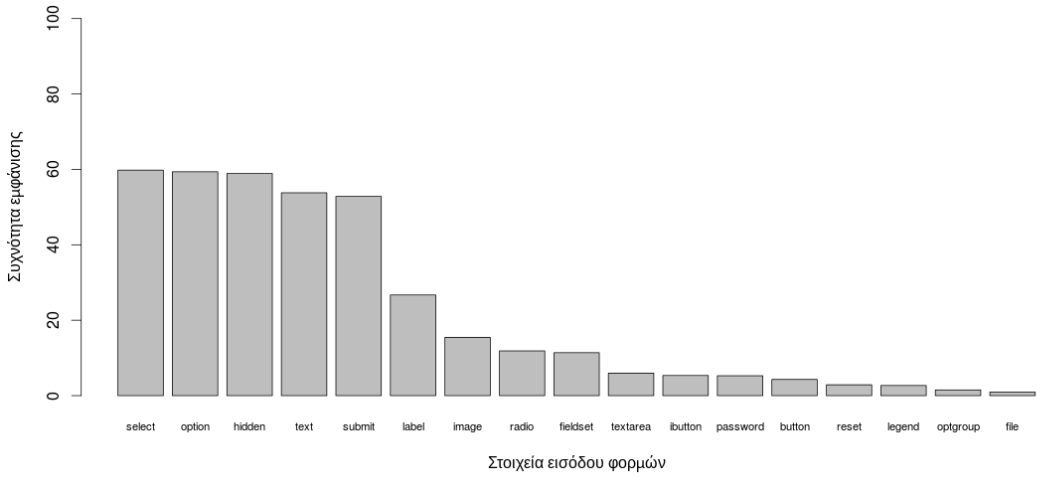
3.6.5.1 Συχνότητα εμφάνισης στοιχείων εισόδου

Στην Εικόνα 3.12, φαίνεται η συχνότητα εμφάνισης των στοιχείων εισόδου στις διεπαφές του δείγματος Yahoo L11. Ένα στοιχείο εισόδου θεωρείται ότι εμφανίζεται σε μια διεπαφή αν το πλήθος του είναι μεγαλύτερο ή ίσο της μονάδας. Από το διάγραμμα καθίσταται σαφές ότι τα στοιχεία select, option, hidden, text και submit εμφανίζονται τουλάχιστον στις μισές διεπαφές, ενώ τα περισσότερο σπάνια είναι τα στοιχεία file, optgroup, legend και reset. Από τη συχνότητα εμφάνισης των στοιχείων είναι δυνατή μια πρώτη υπόθεση σχετικά με τους πιθανούς συσχετισμούς που μπορούν να εξαχθούν από αυτό το δείγμα. Έτσι βάσει της συχνότητας εμφάνισής τους και μόνο, είναι εύλογη η υπόθεση ότι στους όποιους συσχετισμούς υπάρχουν μεταξύ των στοιχείων εισόδου, είναι πιθανότερο να περιλαμβάνονται σαν συνιστώσες, τα στοιχεία εκείνα τα οποία έχουν μεγάλη συχνότητα εμφάνισης (πλήθος μεγαλύτερο της μονάδας) ή/και απουσίας (πλήθος ίσο με το μηδέν).

3.6.5.2 Κεντρική τάση και διασπορά στοιχείων φορμών

Στο διάγραμμα της Εικόνα 3.13 φαίνονται τα θηκογράμματα που αφορούν το πλήθος κάθε στοιχείου εισόδου ανά διεπαφή περιορισμένα μέχρι την τιμή 30. Όπως είναι ορατό, τα τεταρτημόρια των περισσότερων από τα στοιχεία αυτά (legend, fieldset, label, textarea, optgroup, button, input button, image, file, reset, radio, password) είναι πολύ κοντά στο μηδέν, κάτι που

σημαίνει ότι όποτε εμφανίζονται σε μια φόρμα, θα είναι ένα ή δύο. Αυτά τα αποτελέσματα καταδεικνύουν ότι σε μια πιθανή διακριτοποίηση των τιμών αυτών, το πλήθος των κατηγοριών αναμένεται να κυμαίνεται πολύ κοντά σε αυτές τις τιμές. Στη συνέχεια, τέσσερα άλλα στοιχεία (label, select, hidden, submit, text) επιδεικνύουν ελαφρώς μεγαλύτερη διασπορά, με ενδιάμεσο και μέσο όρο να κυμαίνεται πολύ κοντά στο 1, ενώ υπάρχουν αρκετές περιπτώσεις που ξεπερνάνε την τιμή 2.



Εικόνα 3.12 - Συχνότητα εμφάνισης στοιχείων εισόδου διεπαφών

Τα στοιχεία label και text παρουσιάζουν παρόμοια διασπορά και αυτό είναι λογικό καθώς συνηθίζουν να εμφανίζονται μαζί. Αντίθετα τα κουμπιά τύπου submit που συνηθίζεται να είναι ένα ανά φόρμα, εμφανίζονται περισσότερα σε φόρμες που αποτελούν οδηγούς για τη διεκπεραίωση διαδικασιών, όπως π.χ. την αγορά ενός προϊόντος. Τέλος το στοιχείο option είναι το μοναδικό στοιχείο που ξεπερνά τα όρια των 30 στοιχείων ανά διεπαφή και αυτό είναι απόλυτα λογικό, αφού οι επιλογές σε ένα στοιχείο select τείνουν να είναι πολλές (ενδιάμεσος 9, μέσος όρος 47, μέγιστη τιμή 8603). Για παράδειγμα, για την εμφάνιση μιας ημερομηνίας χρειάζονται τουλάχιστον 43 επιλογές (31 για τις ημέρες και 12 για τους μήνες).

3.6.5.3 Συσχετίσεις στοιχείων εισόδου

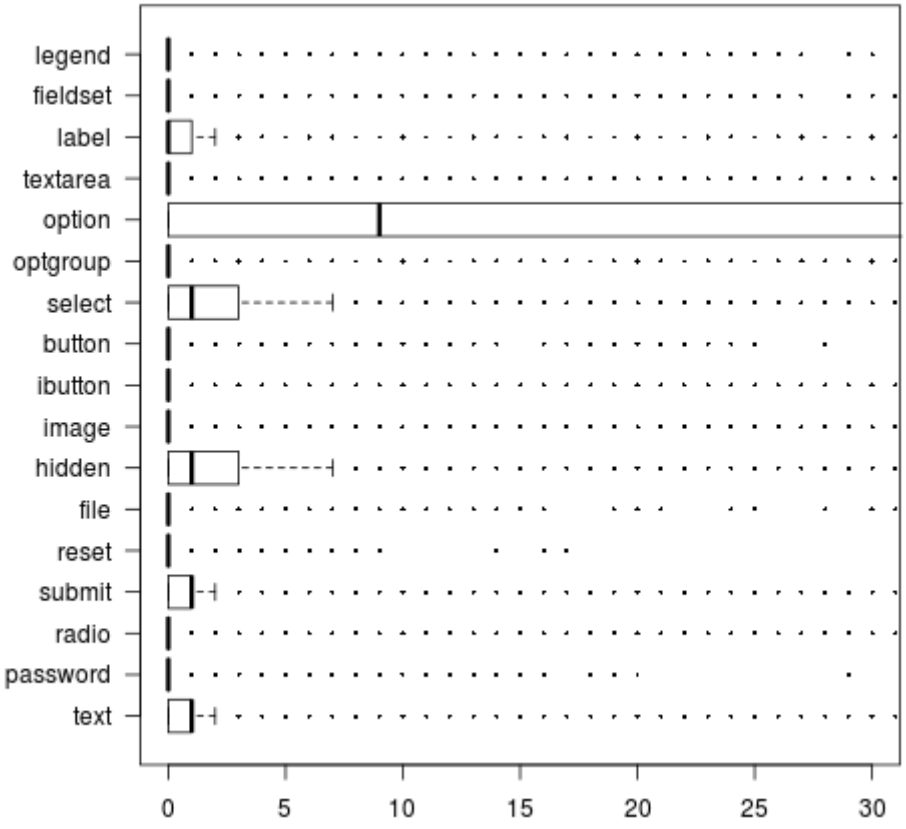
Στην Εικόνα 3.14 φαίνεται το διάγραμμα συσχετίσεων των στοιχείων εισόδου. Για τη δημιουργία του χρησιμοποιήθηκε ο συντελεστής γραμμικής συσχέτισης Pearson που ορίζεται ως:

$$r = \frac{s_{xy}}{s_x s_y} \tag{3.2}$$

όπου:

$$s_{xy} = Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{3.3}$$

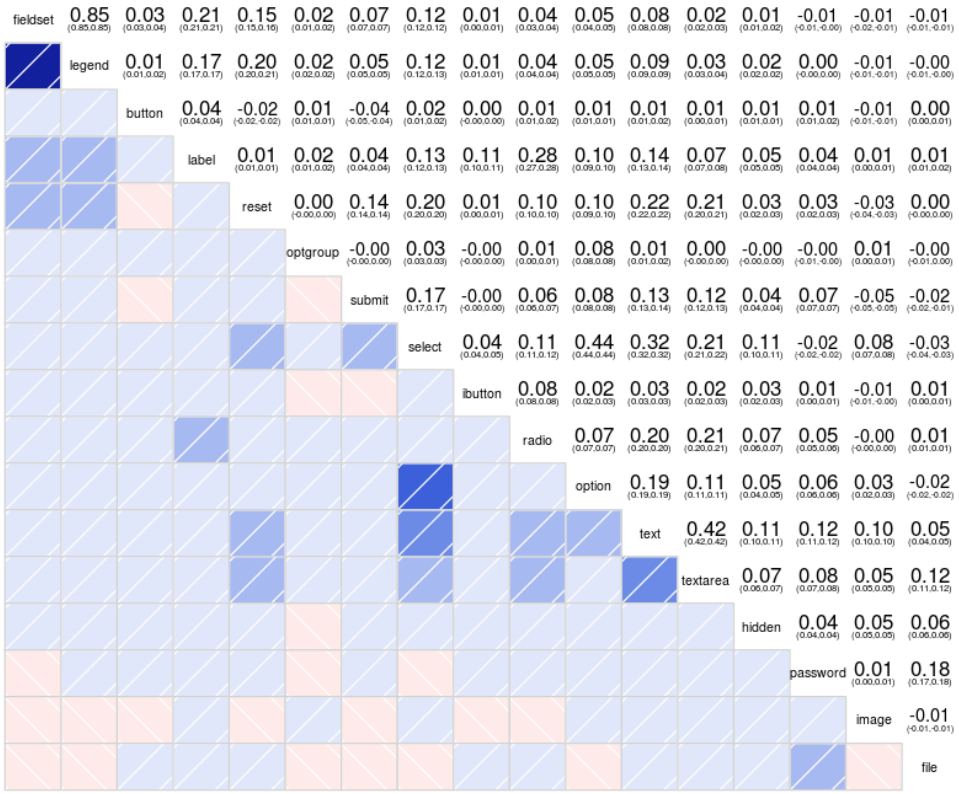
$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})}, s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})}, \tag{3.4}$$



Εικόνα 3.13 - Θηκογράμματα πλήθους στοιχείων εισόδου ανά διεπαφή

Στο διάγραμμα η διακύμανση του μπλε και κόκκινου χρώματος προσδιορίζει το μέγεθος συσχέτισης του κάθε ζευγαριού μεταβλητών. Πιο συγκεκριμένα το έντονο μπλε αντιστοιχεί σε δυνατή θετική συσχέτιση, το έντονο κόκκινο σε δυνατή αρνητική και οτιδήποτε ανάμεσά τους σε πιο χαλαρή ή μηδενική συσχέτιση. Στη συντριπτική πλειονότητα οι συσχετίσεις των στοιχείων εισόδου είναι πολύ κοντά στο μηδέν, κάτι που σημαίνει ότι δεν υπάρχει γραμμική συσχέτιση μεταξύ των μεταβλητών, άρα δεν είναι δυνατή η πρόβλεψη της τιμής ενός στοιχείου από την τιμή του άλλου. Ωστόσο δυο από τα ζευγάρια στοιχείων ξεχωρίζουν: α) το ζευγάρι fieldset-legend και β) το ζευγάρι

select-option. Και στα δυο ζευγάρια τα στοιχεία είναι λογικό να συσχετίζονται καθότι υπάρχει μια φυσική αντιστοιχία ένα-προς-ένα για το πρώτο ζευγάρι (συνήθως κάθε σύνολο πεδίων έχει και μια επικεφαλίδα) ενώ στο δεύτερο ζευγάρι τα στοιχεία options ορίζονται μέσα σε στοιχεία select, οπότε όσο περισσότερα selects υπάρχουν σε μια φόρμα τόσο περισσότερα θα είναι και τα options.



Εικόνα 3.14 - Πίνακας συσχετίσεων στοιχείων εισόδου

3.6.5.4 Ανάλυση κειμένου διεπαφών

Πέραν του πλήθους των στοιχείων ανά φόρμα, είναι δυνατόν ένας αλγόριθμος κατηγοριοποίησης να εκμεταλλευτεί το κείμενο εντός της. Στην περίπτωση αυτή ωστόσο, είτε θα πρέπει να δημιουργηθούν πολλαπλά μοντέλα, ένα για κάθε γλώσσα που πρέπει να υποστηρίζει το σύστημα, είτε θα δημιουργηθεί ένα μοντέλο που θα εμπεριέχει όλες τις περιπτώσεις γλωσσών από τις οποίες έχει εκπαιδευτεί. Στη δεύτερη περίπτωση ο χώρος ιδιοτήτων και κατ' επέκταση ο χρόνος εκπαίδευσης αυξάνεται δραματικά. Στόχος της μελέτης αυτής λοιπόν είναι να εξεταστεί αν υπάρχουν λέξεις κοινές σε όλες τις γλώσσες του δείγματος Yahoo L11. Έτσι λοιπόν από κάθε διεπαφή, εξήχθη το κείμενό της, οι τιμές των ιδιοτήτων της ετικέτας της και οι τιμές των πεδίων εισόδου της, όπου υπήρχαν. Στον Πίνακας 3.6 φαίνονται οι 20 πιο συχνές λέξεις από το περιεχόμενο όλων των διεπαφών του δείγματος. Από τις λέξεις αυτές 4 ξεχωρίζουν τόσο ως προς την ισόποση κατανομή τους στις γλώσσες του δείγματος. Οι λέξεις αυτές είναι οι post, get, search και mail. Οι δυο πρώτες αποτελούν

τιμές της ιδιότητας method της ετικέτας form των διεπαφών, ενώ οι άλλες δύο, προέρχονται από το κείμενο των διεπαφών. Κατά συνέπεια είναι δυνατή η προσθήκη τριών δυαδικών ιδιοτήτων που θα λαμβάνονται υπόψη κατά την εκπαίδευση. Αυτές είναι α) η μέθοδος υποβολής της διεπαφής με τιμές τις "get" και "post", β) η ύπαρξη της λέξης "search" με τιμές τις "YES" (αν υπάρχει) και "NO" (αν δεν υπάρχει) και γ) η ύπαρξη της λέξης "mail" με τιμές "YES" και "NO".

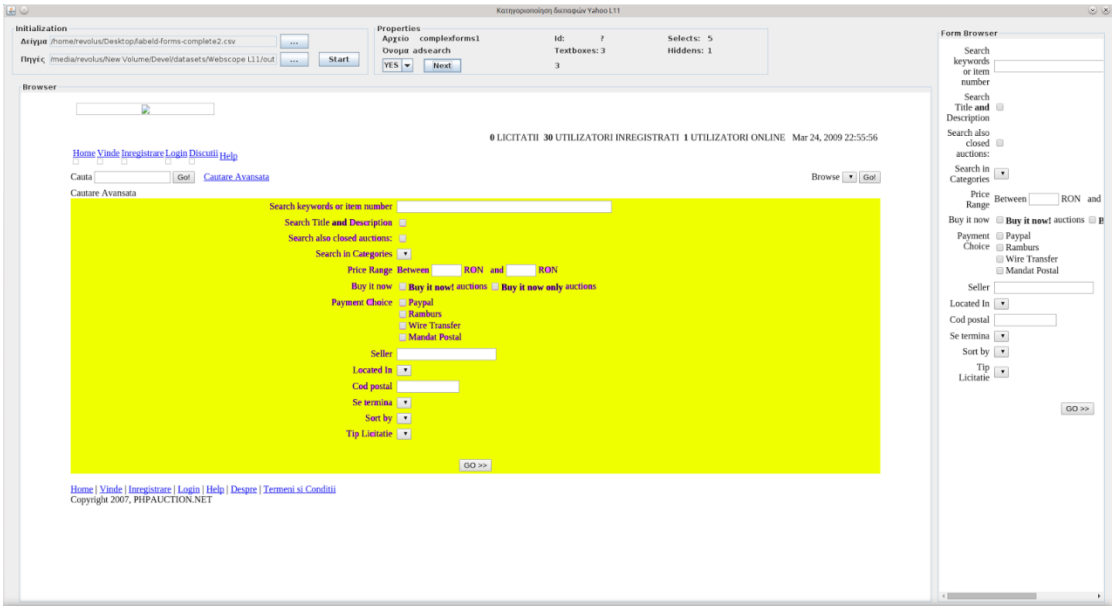
Πίνακας 3.6 - Οι 20 πιο συχνοί όροι στις διεπαφές του δείγματος Yahoo L11

Όρος		**Γλώσσα**	
	Συχνότητα	**Αγγλικά**	**Άλλη**
post	*0.5266863087*	*0.4129283215*	*0.5870716785*
get	*0.4016448909*	*0.5776789714*	*0.4223210286*
search	*0.355174209*	*0.6608091427*	*0.3391908573*
http	0.3217910446	0.469953487	0.530046513
php	0.2820571455	0.405931038	0.594068962
com	0.2355921675	0.5401221763	0.4598778237
all	0.1783056618	0.7319387958	0.2680612042
www	0.1732683715	0.480187101	0.519812899
new	0.1379396707	0.7583368105	0.2416631895
amp	0.1373298758	0.7839984444	0.2160015556
mail	*0.1270273998*	*0.430344791*	*0.569655209*
html	0.1148413522	0.6357013367	0.3642986633
go	0.1122155003	0.8424029444	0.1575970556
citi	0.1065010057	0.9538363788	0.0461636212
de	0.1002368659	0.1589454163	0.8410545837
state	0.0946836351	0.9730310681	0.0269689319
select	0.0922719375	0.8911535954	0.1088464046
type	0.0862022476	0.8889964961	0.1110035039
us	0.0799640345	0.9151881825	0.0848118175
from	0.0794919101	0.9317032123	0.0682967877

3.6.6 Κατηγοριοποίηση διεπαφών συνόλου δεδομένων Yahoo L11

Η δημιουργία ενός συνόλου εκπαίδευσης πάνω στο οποίο ένας κατηγοριοποιητής θα βασιστεί για την κατασκευή ενός μοντέλου κατηγοριοποίησης χρειάζεται ανθρώπινη παρέμβαση, η οποία τις περισσότερες φορές περιλαμβάνει απλώς τον προσδιορισμό της κατηγορίας κάθε στιγμιότυπου στο σύνολο δεδομένων. Στο δείγμα Yahoo L11 χρειάζεται να προσδιοριστεί το είδος της λειτουργικότητας κάθε φόρμας και πιο συγκεκριμένα αν είναι διεπαφή αναζήτησης ή όχι. Η διαδικασία αυτή είναι μια επίπονη και χρονοβόρος, καθώς χρειάζεται η παρατήρηση κάθε φόρμας ξεχωριστά [89]. Για το λόγο αυτό δεν εξετάστηκε ολόκληρο το σύνολο διεπαφών του Yahoo L11, αλλά ένα μικρό τυχαίο δείγμα. Επιπρόσθετα για το σκοπό αυτό δημιουργήθηκε μια ειδική εφαρμογή για τη διευκόλυνση της χειροκίνητης κατηγοριοποίησης, η οποία προσπελαύνει μια προς μια τις φόρμες του δείγματος που προέκυψαν με τη μεθοδολογία της παραγράφου 3.6.3 και την εμφανίζει τονισμένη μέσα στην ιστοσελίδα της. Έτσι ένας ειδικός μπορεί εύκολα να αναγνωρίσει αν η φόρμα αφορά την αναζήτηση δεδομένων ή όχι και να ορίσει την κατηγορία άμεσα. Μια άποψη της

συγκεκριμένης εφαρμογής φαίνεται στην εικόνα 3.15. Έτσι με το τυχαίο δείγμα που παράχθηκε και την εφαρμογή χειροκίνητης κατηγοριοποίησης που δημιουργήθηκε, η όλη διαδικασία της κατηγοριοποίησης απλοποιήθηκε αρκετά.



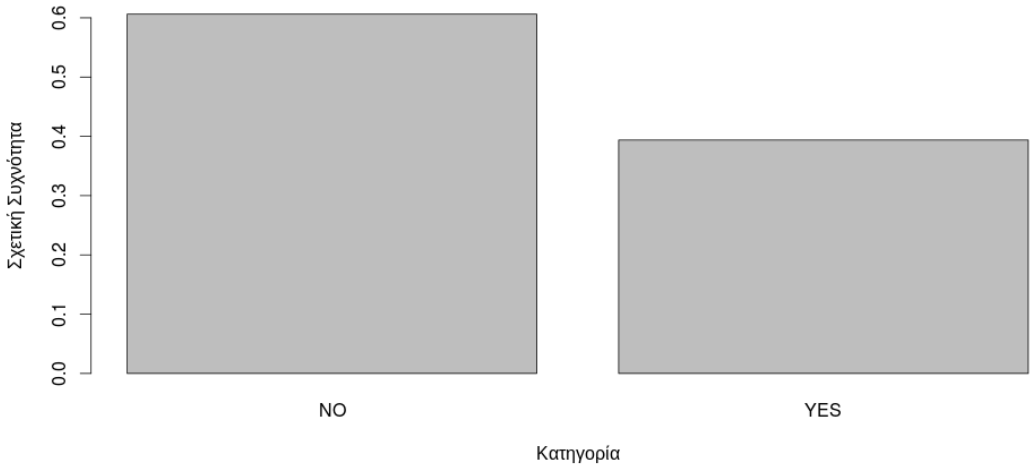
Εικόνα 3.15 - Άποψη της ειδικής εφαρμογής που δημιουργήθηκε για την κατηγοριοποίηση των διεπαφών του Yahoo L11

3.6.6.1 Κατανομή συχνοτήτων κατηγοριών

Στην Εικόνα 3.16 φαίνεται η κατανομή συχνοτήτων του είδους της λειτουργικότητας των φορμών του δείγματος επί του συνόλου των διεπαφών του Yahoo L11 (YES: διεπαφή αναζήτησης, NO: μη διεπαφή αναζήτησης. Στο δείγμα μεγέθους 1803 φορμών, το ποσοστό των διεπαφών αναζήτησης είναι 39,3%.

3.7 Επαγωγή κανόνων για την κατηγοριοποίηση διεπαφών κατά λειτουργία

Στην παράγραφο αυτή εξετάζεται η χρήση της προσέγγισης που παρουσιάστηκε στο Κεφάλαιο 2 για την επαγωγή κανόνων που θα χρησιμεύσουν στην «προ-ερωτηματική» κατηγοριοποίηση διεπαφών ως προς τη λειτουργία τους. Οι επιδόσεις της συγκεκριμένης προσέγγισης για διάφορες τιμές των παραμέτρων της, μελετώνται για να γίνει η βέλτιστη επιλογή τους και στη συνέχεια η προσέγγιση συγκρίνεται με μια σειρά γνωστών αλγορίθμων κατηγοριοποίησης αλλά και έτοιμων κανόνων που ωστόσο κατασκευάστηκαν χειροκίνητα από τους συγγραφείς των αντίστοιχων εργασιών.



Εικόνα 3.16 - Κατανομή συχνότητας κατηγοριών

Ως δεδομένα εκπαίδευσης χρησιμοποιούνται δυο σύνολα δεδομένων: α) το δείγμα του Yahoo L11 που παρουσιάστηκε στην παράγραφο 3.6.6 και β) ένα δεύτερο δείγμα που βασίζεται στο σύνολο TEL-8 από το αποθετήριο UIUC [100]. Το δεύτερο αυτό σύνολο, περιέχει διεπαφές αναζήτησης από 8 θεματικές κατηγορίες, μαζί με τις σελίδες στις οποίες αυτές περιέχονται. Το TEL-8 έχει χρησιμοποιηθεί σε πολλαπλές εργασίες, οι περισσότερες εκ των οποίων ωστόσο, στόχευαν στη θεματική κατηγοριοποίηση των διεπαφών. Για το λόγο αυτό, έγινε χειροκίνητη κατηγοριοποίησή τους ως προς τη λειτουργία τους. Η περιγραφή των δυο συνόλων φαίνεται στον Πίνακας 3.7

Τα δεδομένα εκπαίδευσης που δίνονται ως είσοδος στους αλγόριθμους αυτούς αποτελούνται από τις εξής ιδιότητες: α) τη συχνότητα εμφάνισης όλων των πιθανών στοιχείων εισόδου όπως περιγράφονται στην προδιαγραφή της HTML, όπως αρχικά είχε αναφερθεί και β) τις δυαδικές ιδιότητες που προέκυψαν από την καταμέτρηση των συχνοτήτων των λέξεων των διεπαφών του συνόλου Yahoo L11 και πιο συγκεκριμένα τη μέθοδο υποβολής ("get" και "post") και την ύπαρξη των λέξεων "search" και "mail" με τιμές τις "YES" και "NO".

Πίνακας 3.7 - Τα σύνολα δεδομένων που χρησιμοποιήθηκαν για εκπαίδευση

Σύνολο Δεδομένων	Πλήθος παραδειγμάτων	Πλήθος Ιδιοτήτων	Πλήθος Κατηγοριών	Κατανομή Κατηγοριών
Yahoo L11	1803	25	2	710(NAI)/1093(OXI)
TEL-8	1114	25	2	450(NAI)/664(OXI)

3.7.1 Εμπειρική εκτίμηση των παραμέτρων προτεινόμενης προσέγγισης

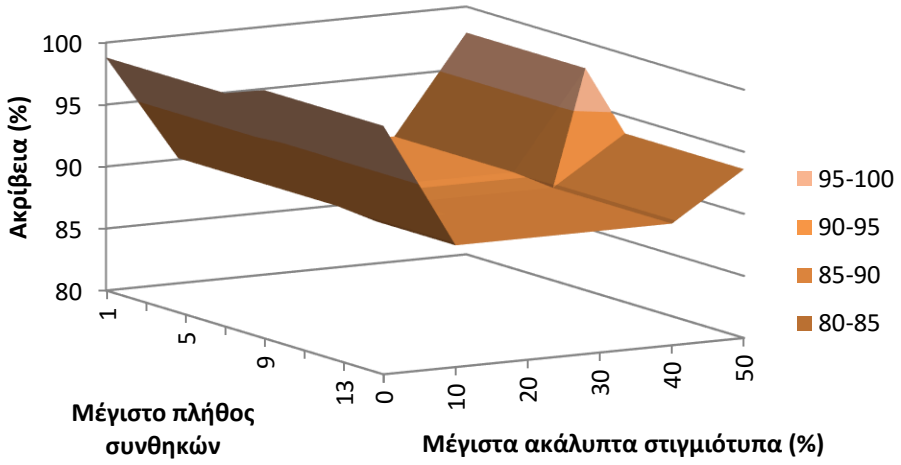
Σε αυτή την παράγραφο γίνεται μια πειραματική μελέτη της ακρίβειας, του πλήθους και του μεγέθους των παραγόμενων κανόνων της προτεινόμενης μεθόδου για διάφορες τιμές των δυο παραμέτρων της. Οι δυο παράμετροι αυτοί είναι α) το μέγιστο πλήθος των στιγμιότυπων τα οποία

θα αφήνουν ακάλυπτα οι κανόνες και β) το μέγιστο πλήθος συνθηκών που μπορεί να υπάρχουν σε έναν κανόνα. Όλα τα πειράματα εκτελέστηκαν στα δυο προαναφερθέντα σύνολα δεδομένων του Πίνακας 3.7, χρησιμοποιώντας εναλλάξ το ένα για σύνολο εκπαίδευσης και το άλλο για σύνολο δοκιμής. Το μέτρο αξιολόγησης της προσέγγισης που επιλέχθηκε είναι η ακρίβεια (accuracy) η οποία σύμφωνα με τη σχέση (3.5), ορίζεται ως το πλήθος των σωστά κατηγοριοποιημένων στιγμιότυπων (ορθώς θετικών TP και ορθώς αρνητικών TN) προς το πλήθος όλων των στιγμιότυπων που εξετάστηκαν. Η ακρίβεια χρησιμοποιείται ως στατιστικό μέτρο για την αξιολόγηση του πόσο καλά η δυαδική κατηγοριοποίηση αναγνωρίζει την κατηγορία ενός στιγμιότυπου A.

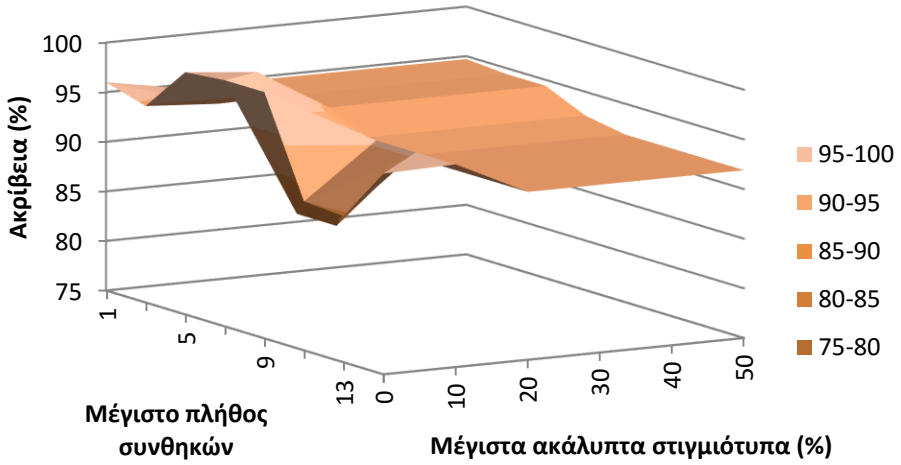
$$Ακρίβεια(A) = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.5}$$

3.7.1.1 Ακρίβεια κατηγοριοποίησης

Στην Εικόνα 3.17 και την Εικόνα 3.18 φαίνεται η ακρίβεια της προτεινόμενης προσέγγισης για διάφορες τιμές των δυο παραμέτρων εισόδου, για τα σύνολα Yahoo L11 και TEL-8.



Εικόνα 3.17 - Ακρίβεια κατηγοριοποίησης προτεινόμενης προσέγγισης με σύνολο εκπαίδευσης το Yahoo L11 και δοκιμής το TEL-8 για διάφορες τιμές των παραμέτρων



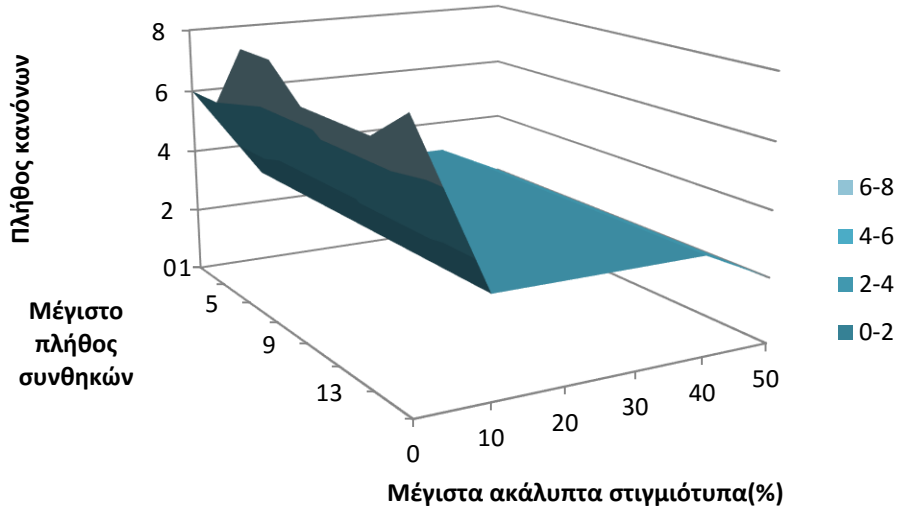
Εικόνα 3.18 - Ακρίβεια κατηγοριοποίησης προτεινόμενης προσέγγισης με σύνολο εκπαίδευσης το TEL-8 και δοκιμής το Yahoo L11 για διάφορες τιμές των παραμέτρων

Στην Εικόνα 3.17 η μέγιστη ακρίβεια παρατηρείται στις ακραίες τιμές της παραμέτρου του ποσοστού ακάλυπτων στιγμιότυπων και πιο συγκεκριμένα στις τιμές 0% και 50%, με ακρίβεια 98% και 97% αντίστοιχα. Αντίθετα, η ακρίβεια παραμένει σταθερή (90%) για όλες τις ενδιάμεσες τιμές. Η επεξήγηση του φαινομένου αυτού φαίνεται στην Εικόνα 3.19. Ο αλγόριθμος στην προσπάθειά του να καλύψει όλο το σύνολο δεδομένων (0% ακάλυπτα παραδείγματα) παράγει το περισσότερο πλήθος κανόνων (6). Όσο λιγότερα παραδείγματα καλύπτει ωστόσο, το πλήθος των κανόνων παραμένει σταθερό (3), μέχρι τη μέγιστη τιμή που λαμβάνει (50%), οπότε και οι κανόνες που παράγονται είναι οι λιγότεροι (2). Αυτό σημαίνει ότι στις ενδιάμεσες τιμές αφενός παράγονταν οι ίδιοι (ή ισοδύναμοι) κανόνες και αφετέρου με την αφαίρεση του ενός, βελτιώθηκε η ακρίβεια.

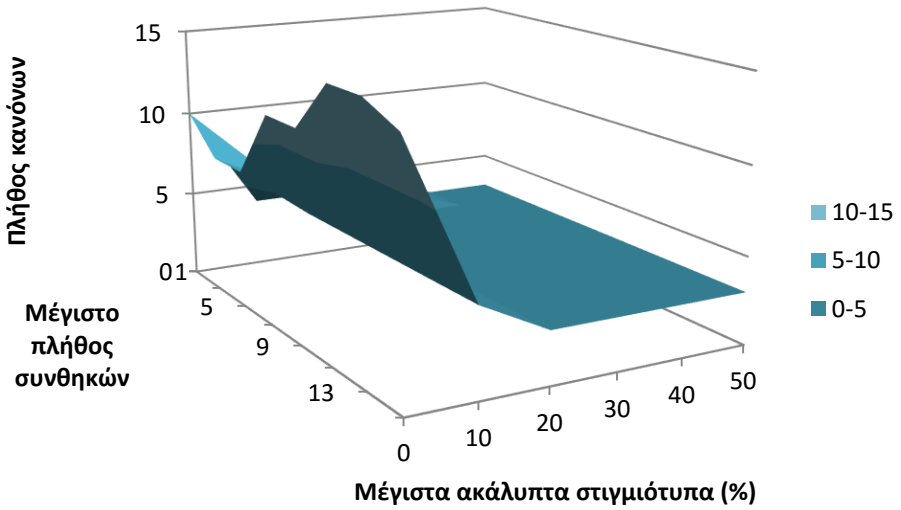
Στην Εικόνα 3.18, δεν παρατηρείται το ίδιο φαινόμενο. Η μέγιστη τιμή της ακρίβειας (98%) παρατηρείται στις χαμηλές τιμές της παραμέτρου των ακάλυπτων στιγμιότυπων (0%-10%). Αυτό σημαίνει ότι οι πρώτοι κανόνες στη λίστα των κανόνων και μόνο αυτοί είναι που διαμορφώνουν την ακρίβεια για όλες τις υπόλοιπες τιμές της παραμέτρου (20%-50%). Από τα δύο αυτά σχήματα μπορεί να εξαχθεί το συμπέρασμα ότι οι βέλτιστες τιμές για τις παραμέτρους κυμαίνονται στο 0%-10% για το ποσοστό ακάλυπτων στιγμιότυπων και 3-6 μέγιστο πλήθος συνθηκών ανά κανόνα. Αυτό μένει να επιβεβαιωθεί και από το πλήθος και το μέγεθος των κανόνων που παράγονται.

3.7.1.2 Πλήθος κανόνων

Στην Εικόνα 3.19 και την Εικόνα 3.20 φαίνεται το πλήθος παραγόμενων κανόνων της προτεινόμενης προσέγγισης για διάφορες τιμές των δυο παραμέτρων εισόδου, για τα σύνολα Yahoo L11 και TEL-8.



Εικόνα 3.19 - Πλήθος κανόνων προτεινόμενης προσέγγισης με σύνολο εκπαίδευσης το Yahoo L11 και δοκιμής το TEL-8 για διάφορες τιμές των παραμέτρων



Εικόνα 3.20 – Πλήθος κανόνων προτεινόμενης προσέγγισης με σύνολο εκπαίδευσης το TEL-8 και δοκιμής το Yahoo L11 για διάφορες τιμές των παραμέτρων

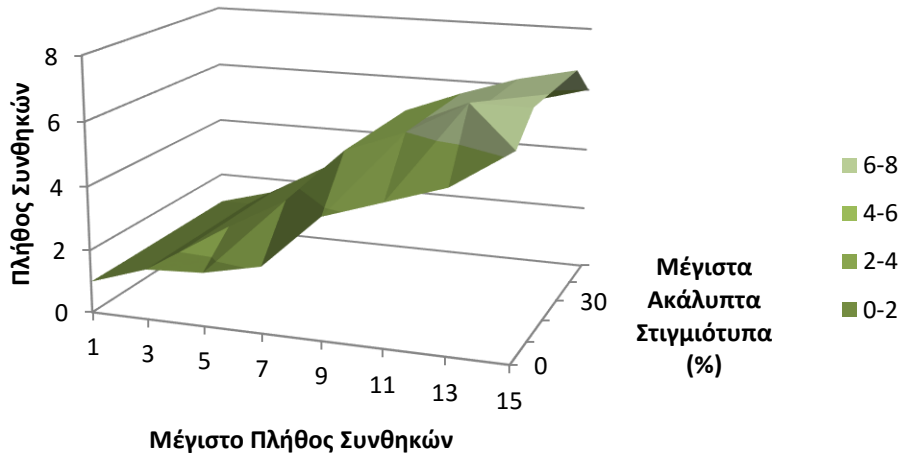
Και στις δυο εικόνες, παρατηρείται η ίδια πτωτική τάση του πλήθους των κανόνων. Όταν το ποσοστό ακάλυπτων στιγμιότυπων είναι 0% προφανώς παράγονται οι περισσότεροι κανόνες στην προσπάθεια κάλυψης όλου του συνόλου δεδομένων. Για τις τιμές 10%-50% στο ποσοστό ακάλυπτων στιγμιότυπων, το πλήθος των κανόνων παραμένει σταθερό, κάτι που σημαίνει ότι οι κανόνες που παράγονται επαρκούν να καλύψουν το 90% του συνόλου. Σχετικά με το μέγιστο πλήθος συνθηκών παρατηρείται ότι για μικρές τιμές (0-5) παράγονται σχετικά λιγότεροι κανόνες (6-7). Έτσι εξάγεται το συμπέρασμα, ότι οι βέλτιστες τιμές για τις δυο παραμέτρους, όσον αφορά το πλήθος των κανόνων

κυμαίνονται από το 0-10% για το ποσοστό ακάλυπτων στιγμιότυπων και 1-5 για το μέγιστο πλήθος συνθηκών ανά κανόνα.

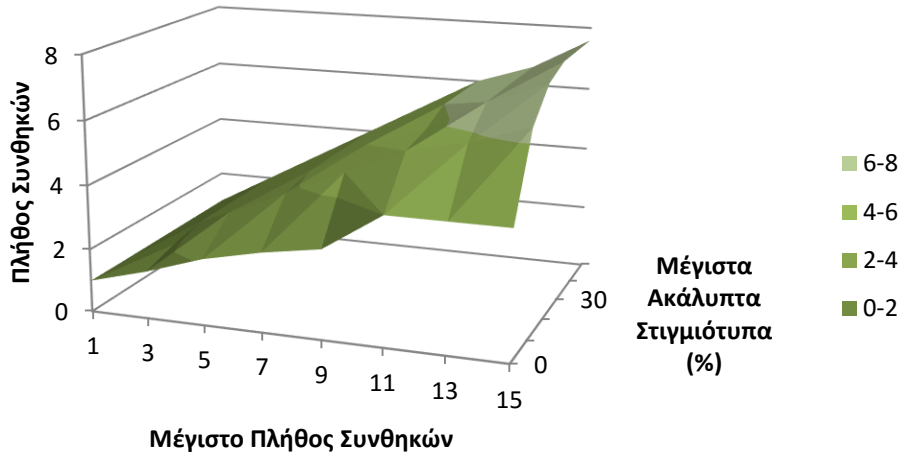
3.7.1.3 Πλήθος συνθηκών ανά κανόνα

Στην Εικόνα 3.17 και την Εικόνα 3.18 φαίνεται το πλήθος των συνθηκών των κανόνων που παράγονται από την προτεινόμενη προσέγγιση για διάφορες τιμές των δυο παραμέτρων εισόδου, για τα σύνολα Yahoo L11 και TEL-8.

Από τις δυο εικόνες είναι ξεκάθαρο ότι το πλήθος των συνθηκών ανά κανόνα εξαρτάται μόνο από την παράμετρο του μέγιστου πλήθους συνθηκών. Ωστόσο επειδή ο αλγόριθμος πραγματοποιεί εξαντλητική αναζήτηση στο χώρο των κανόνων με τις καλύτερες συνθήκες που έχει εντοπίσει σε μια δεδομένη επανάληψη, γενικά δεν είναι υποχρεωτικό να αυξάνεται το πλήθος των συνθηκών, αν η προσθήκη τους χειροτερεύει την ποιότητα των κανόνων. Αυτό που εξάγεται σαν συμπέρασμα από τα αποτελέσματα της μελέτης, είναι ότι ουσιαστικά υπάρχουν αρκετές συσχετιζόμενες συνθήκες που οδηγούν σε ποιοτικούς κανόνες. Ωστόσο επειδή με ένα μικρό πλήθος συνθηκών αποφεύγεται το φαινόμενο της υπερπροσαρμογής, είναι προτιμότερο, το μέγιστο πλήθος τους να τηρείται χαμηλό.



Εικόνα 3.21 - Πλήθος συνθηκών ανά κανόνα προτεινόμενης προσέγγισης με σύνολο εκπαίδευσης το Yahoo L11 και δοκιμής το TEL-8 για διάφορες τιμές των παραμέτρων



Εικόνα 3.22 - Πλήθος συνθηκών ανά κανόνα για την προτεινόμενη προσέγγιση με σύνολο εκπαίδευσης το TEL-8 και δοκιμής το Yahoo L11 για διάφορες τιμές των παραμέτρων

Από τις παραπάνω παραγράφους, οι παράμετροι επιλέγονται ως εξής:

- Μέγιστο ποσοστό ακάλυπτων παραδειγμάτων: 5%
- Μέγιστο πλήθος συνθηκών 5

3.8 Σύγκριση προτεινόμενης προσέγγισης με άλλες προσεγγίσεις

Οι εναλλακτικές προσεγγίσεις που εξετάστηκαν στα πλαίσια της αξιολόγησης της προτεινόμενης προσέγγισης, περιλαμβάνουν δυο χειροκίνητα κατασκευασμένους κανόνες που έχουν προταθεί στις εργασίες [95] και [108]. Η λίστα κανόνων της προσέγγισης [95] είναι η ακόλουθη:

- Μια διεπαφή είναι διεπαφή αναζήτησης αν η μέθοδος υποβολής της είναι η GET και αν δεν παρουσιάζει πεδίο εισόδου τύπου κωδικού.
- Μια διεπαφή είναι διεπαφή αναζήτησης αν η μέθοδος υποβολής της είναι η GET και περιέχει οπουδήποτε τη λέξη "search".
- Μια διεπαφή είναι διεπαφή αναζήτησης αν δεν παρουσιάζει πεδίο εισόδου τύπου κωδικού και περιέχει οπουδήποτε τη λέξη "search".
- Μια διεπαφή είναι διεπαφή αναζήτησης αν η μέθοδος υποβολής της είναι η GET, αν δεν παρουσιάζει πεδίο εισόδου τύπου κωδικού και αν περιέχει οπουδήποτε τη λέξη "search".

Η διάταξη της παραπάνω λίστας θα μπορούσε να είναι όπως φαίνεται, ή να τοποθετηθεί ο τελευταίος κανόνας στην αρχή, όντας ο πιο ειδικός αφού έχει περισσότερες συνθήκες και κατά συνέπεια θα καλύπτει λιγότερα παραδείγματα. Επίσης, είναι προφανές πως οι παραπάνω κανόνες αποτελούνται μόνο από συμπεράσματα που κατηγοριοποιούν τις διεπαφές για τις οποίες ισχύουν ως

διεπαφές αναζήτησης. Θα ήταν λογική συνέπεια λοιπόν να υποτεθεί ότι όσες διεπαφές δεν περιγράφονται από τους παραπάνω κανόνες δεν είναι διεπαφές αναζήτησης. Άρα μπορεί να προστεθεί ένας τελευταίος κανόνας που θα χαρακτηρίζει κάθε διεπαφή που φτάνει σε αυτόν ως μη αναζήτησης.

Από την άλλη η λίστα κανόνων της προσέγγισης [108] είναι η ακόλουθη:

- Αν μια διεπαφή περιέχει ένα στοιχείο εισόδου τύπου password, τότε δεν είναι διεπαφή αναζήτησης.
- Αν μια διεπαφή περιέχει ένα στοιχείο εισόδου τύπου mail, τότε δεν είναι διεπαφή αναζήτησης.
- Αν μια διεπαφή περιέχει ένα στοιχείο εισόδου τύπου button και οποιεσδήποτε από τις ιδιότητές του περιέχουν μια από τις λέξεις "search", "find", "go", "buy" τότε είναι διεπαφή αναζήτησης.
- Αν μια διεπαφή περιέχει ένα στοιχείο εισόδου τύπου image και οποιεσδήποτε από τις ιδιότητές του περιέχουν μια από τις λέξεις "search", "find", "go", "buy" τότε είναι διεπαφή αναζήτησης.
- Αν μια διεπαφή περιέχει ένα στοιχείο εισόδου τύπου submit και οποιεσδήποτε από τις ιδιότητές του περιέχουν μια από τις λέξεις "search", "find", "go", "buy" τότε είναι διεπαφή αναζήτησης.
- Αν μια διεπαφή περιέχει ένα στοιχείο τύπου button και οποιεσδήποτε από τις ιδιότητές του περιέχουν μια από τις λέξεις "search", "find", "go", "buy" τότε είναι διεπαφή αναζήτησης.
- Αν μια διεπαφή περιέχει ένα στοιχείο εισόδου τύπου img και οποιεσδήποτε από τις ιδιότητές του περιέχουν μια από τις λέξεις "search", "find", "go", "buy" τότε είναι διεπαφή αναζήτησης.

Στην παραπάνω λίστα είναι εμφανές ότι οι συγγραφείς περιλαμβάνουν κανόνες που καλύπτουν και τις δυο κατηγορίες υπό εξέταση, δηλαδή τις διεπαφές αναζήτησης και τις υπόλοιπες. Κατά συνέπεια η προσθήκη ενός επιπλέον κανόνα όπως στην περίπτωση της εργασίας [95] δεν έχει νόημα, γιατί δεν υπάρχει κάποιο κριτήριο για να επιλεχθεί η μία έναντι της άλλης.

Για το συγκεκριμένο πείραμα, πέραν των δύο παραπάνω λιστών κανόνων, χρησιμοποιήθηκαν οι εξής αλγόριθμοι κατηγοριοποίησης: J48, Multilayer Perceptron, SOM, Naïve Bayes όπως είχαν χρησιμοποιηθεί στην εργασία [96] και επιπρόσθετα χρησιμοποιήθηκαν δυο αλγόριθμοι επαγωγής κανόνων που μπορούν να χειριστούν αριθμητικά δεδομένα: ο JRip και ο Decision Table. Για τους αλγόριθμους αυτούς χρησιμοποιήθηκαν οι υλοποιήσεις που παρέχονται στο πακέτο WEKA [61] με τις προεπιλεγμένες τους παραμέτρους. Στις περιπτώσεις που το σύνολο εκπαίδευσης ήταν το ίδιο με το σύνολο δοκιμής, εφαρμόστηκε δεκαπλή σταυρωτή επικύρωση (10-fold cross validation). Για τις

λίστες κανόνων των προσεγγίσεων [95] και [108] τα σύνολα προφανώς χρησιμοποιήθηκαν μόνο για δοκιμή. Τα αποτελέσματα των μετρήσεων φαίνονται στον Πίνακας 3.8. Οι καλύτερες αποδόσεις ανά περίπτωση και αλγόριθμο εμφανίζονται με εντονότερο χρώμα.

Πίνακας 3.8 - Ακρίβεια κατηγοριοποίησης στα δείγματα των συνόλων δεδομένων Yahoo L11 και TEL-8

Εκπ./Δοκιμή	Προτ.	J48	M. P.	SOM	Bayes	JRip	D.T.	[95]	[108]
YL11/YL11	**99.1**	90.6	87.9	86.6	84.1	90.3	87.5	94.0	75.0
TEL8/TEL8	**97.8**	87.6	85.5	82.9	71.7	88.1	88.5	93.0	50,0
YL11/TEL8	**98.1**	83.7	72.8	72.6	75.3	84.1	84.5	93.0	50,0
TEL8/YL11	**98.7**	82.5	74.2	83.9	72.1	83.7	74.3	94.0	75.0

Από τα αποτελέσματα του Πίνακας 3.8, μια πρώτη παρατήρηση αφορά την ακρίβεια κατηγοριοποίησης της προτεινόμενης προσέγγισης (πρώτη στήλη) που αποτελεί την καλύτερη όλων των εναλλακτικών προσεγγίσεων. Παρά την πτώση στην ακρίβεια κατηγοριοποίησης σε όλες τις υπόλοιπες προσεγγίσεις όταν γίνεται εκπαίδευση ή δοκιμή στο σύνολο TEL-8, η ακρίβεια κατηγοριοποίησης της προτεινόμενης προσέγγισης παραμένει υψηλή κάτι που σημαίνει ότι οι κανόνες γενικεύουν ικανοποιητικά και μπορούν να εφαρμοστούν και σε άλλα σύνολα δεδομένων. Μια δεύτερη παρατήρηση αφορά την επίδοση της λίστας κανόνων της εργασίας [95], που παρά την απλότητά της είναι αρκετά ακριβής. Αυτό δεν ισχύει για τη δεύτερη λίστα κανόνων της εργασίας [108]. Για τη λίστα αυτή ωστόσο κάτι τέτοιο μπορεί να αποδοθεί στο ότι δεν στοχεύει πρωταρχικά στην κατηγοριοποίηση διεπαφών, αλλά στο φιλτράρισμα των διεπαφών που δεν προορίζονται για αναζήτηση και την είσοδό τους σε κάποιο αλγόριθμο κατηγοριοποίησης, για τη βελτίωση των επιδόσεών του.

Οι κανόνες που προέκυψαν από την εκπαίδευση του προτεινόμενου αλγόριθμου στο σύνολο TEL-8 είναι οι εξής:

- Αν δεν υπάρχουν πεδία εισόδου τύπου text και δεν υπάρχουν πεδία εισόδου τύπου select και δεν υπάρχει πεδίο εισόδου τύπου submit τότε η διεπαφή δεν είναι διεπαφή αναζήτησης.
- Αλλιώς αν δεν υπάρχει πεδίο εισόδου τύπου κωδικού και δεν υπάρχει η λέξη mail και υπάρχει η λέξη search τότε αυτή είναι διεπαφή αναζήτησης.
- Αλλιώς η διεπαφή δεν είναι διεπαφή αναζήτησης.

Οι κανόνες που προέκυψαν από την εκπαίδευση του αλγόριθμου στο σύνολο Yahoo L11 είναι οι εξής:

- Αν δεν υπάρχει η λέξη search στη διεπαφή τότε αυτή δεν είναι διεπαφή αναζήτησης
- Αλλιώς αν δεν υπάρχει πεδίο κωδικού και δεν υπάρχουν πεδία εισόδου τύπου file και δεν υπάρχει η λέξη mail στη διεπαφή, τότε αυτή είναι διεπαφή αναζήτησης.

- Αλλιώς αν η μέθοδος υποβολής είναι POST τότε αυτή δεν είναι διεπαφή αναζήτησης
- Αλλιώς αν η μέθοδος υποβολής είναι GET τότε είναι διεπαφή αναζήτησης
- Αλλιώς δεν είναι διεπαφή αναζήτησης

3.9 Συμπεράσματα

Στο κεφάλαιο αυτό, μετά από μια εισαγωγή στο πρόβλημα της αναγνώρισης των διεπαφών αναζήτησης στον Παγκόσμιο Ιστό, έγινε μια ανάλυση του συνόλου δεδομένων Yahoo L11 με τεχνικές ανάλυσης Μεγάλων Δεδομένων. Μια τέτοια προσέγγιση είναι εφαρμόσιμη στην περίπτωση των μηχανών αναζήτησης γενικού σκοπού, όπου συνηθίζεται οι ιστοσελίδες να προσκομίζονται αδιακρίτως σε μια κεντρική τοποθεσία και να αναλύονται εκεί σε δεύτερο βήμα. Έπειτα παρουσιάζονται χρήσιμα συμπεράσματα σχετικά με τις διεπαφές που εξήχθησαν από το σύνολο αυτό που αφορούν τις κατανομές συχνοτήτων εμφάνισης των διαφόρων πεδίων εισόδου τους, τις μεταξύ τους συσχετίσεις, αλλά και τα κειμενικά χαρακτηριστικά που μπορούν να χρησιμοποιηθούν σε μια κατηγοριοποίηση ανεξαρτήτως της γλώσσας στην οποία είναι γραμμένη η διεπαφή.

Πιο συγκεκριμένα στο σύνολο δεδομένων Yahoo L11, λαμβάνοντας υπόψη τους κορυφαίους χώρους ονομάτων βρέθηκε ότι υπάρχει μια σχετικά καλή αντιπροσώπευση του Παγκόσμιου Ιστού, ενώ ταυτόχρονα το δείγμα δεν στοχεύει κάποιες συγκεκριμένες θεματικές κατηγορίες. Επίσης υπάρχει ποικιλία στις εμφανιζόμενες φυσικές γλώσσες, κάτι που σημαίνει ότι το δείγμα μπορεί να αποτελέσει σημείο έναρξης για μελέτη του Κρυμμένου Ιστού μιας συγκεκριμένης χώρας. Στο δείγμα φαίνεται ότι η πλειονότητα των ιστοσελίδων περιέχουν μια ή δυο διεπαφές, ενώ ιδιαίτερα σπάνιες είναι οι σελίδες με πάνω από έξι διεπαφές. Σχετικά με τα στοιχεία εισόδου, τα στοιχεία select, option, hidden, text και submit εμφανίζονται τουλάχιστον στις μισές διεπαφές, ενώ τα περισσότερο σπάνια είναι τα στοιχεία file, optgroup, legend και reset. Σύμφωνα με το σύνολο δεδομένων Yahoo L11, δεν υπάρχει γραμμική συσχέτιση μεταξύ των πεδίων εκτός από τα ζευγάρια πεδίων fieldset-legend και select-option που εμφανίζονται πάντα μαζί σε μια διεπαφή. Αυτό σημαίνει ότι δεν είναι δυνατή η πρόβλεψη της τιμής ενός στοιχείου εισόδου από την τιμή του άλλου. Τέλος από την καταμέτρηση των λέξεων στις διεπαφές φάνηκε ότι οι λέξεις get και post που αποτελούν τιμές της ιδιότητας method μιας διεπαφής, καθώς και οι λέξεις search και mail εμφανίζονται στις διεπαφές ανεξάρτητα από τη φυσική γλώσσα στην οποία είναι γραμμένη. Αυτό σημαίνει ότι μπορούν να χρησιμοποιηθούν σαν ιδιότητες για τη διαδικασία της κατηγοριοποίησης ως προς τη λειτουργία των διεπαφών.

Στη συνέχεια χρησιμοποιείται η προσέγγιση που παρουσιάστηκε στο Κεφάλαιο 2 για την επαγωγή κανόνων κατηγοριοποίησης πάνω σε ένα δείγμα από τις διεπαφές που εξήχθησαν από το σύνολο Yahoo L11 και ένα δεύτερο δείγμα από το σύνολο δεδομένων TEL-8. Οι ιδιότητες που χρησιμοποιήθηκαν προέκυψαν από την ανάλυση του Yahoo L11 με μέριμνα να είναι ανεξάρτητες

της γλώσσας στην οποία είναι γραμμένες οι διεπαφές. Μετά από έναν εμπειρικό προσδιορισμό των παραμέτρων εισόδου, έγινε μια πειραματική μελέτη για την επίδοση τόσο της προτεινόμενης προσέγγισης όσο και άλλων προσεγγίσεων που έχουν προταθεί στη βιβλιογραφία και τα αποτελέσματα δείχνουν ότι η ακρίβεια κατηγοριοποίησης των κανόνων που παρήχθησαν από την προτεινόμενη προσέγγιση υπερέχουν των υπόλοιπων. Αυτό οφείλεται αφενός στην εξαντλητική αναζήτηση που πραγματοποιεί η προτεινόμενη προσέγγιση για την εύρεση του καλύτερου κανόνα σε κάθε επανάληψη, αλλά αφετέρου και στην επιλογή των βέλτιστων παραμέτρων εισόδου για το συγκεκριμένο πρόβλημα. Οι κανόνες που παρήχθησαν είναι σύντομοι σε μήκος και αρκετά γενικοί για να μπορούν να εφαρμοστούν επιτυχώς για την κατηγοριοποίηση διεπαφών ως προς τη λειτουργία τους.

Κεφάλαιο 4

Αναζήτηση πληροφορίας στον Παγκόσμιο Ιστό με έναν αλγόριθμο αποικίας μυρμηγκιών

4.1 Εισαγωγή

Οι διεργασίες που παρατηρούσε ο άνθρωπος στη φύση γύρω του, αποτελούσαν ανέκαθεν έμπνευση για την επίλυση προβλημάτων που αντιμετώπιζε στην καθημερινότητά του. Ωστόσο, μόλις πρόσφατα οι ερευνητές κατάφεραν να μοντελοποιήσουν, να εξομοιώσουν και να εφαρμόσουν τεχνικές εμπνευσμένες από τη φύση σε μια πληθώρα διαφορετικών πεδίων όπως τη μηχανική, την οικονομία και την ιατρική. Ήταν θέμα χρόνου έως ότου οι τεχνικές αυτές να εφαρμόζονταν και στην Επιστήμη των Υπολογιστών.

Στο κεφάλαιο αυτό προτείνεται μια νέα μεθοδολογία αναζήτησης πληροφορίας στον Παγκόσμιο Ιστό που στηρίζεται σε μια από τις πιο γνωστές τεχνικές εμπνευσμένες από τη φύση: της αποικίας μυρμηγκιών. Πιο συγκεκριμένα, προτείνεται μια προσέγγιση με το όνομα Ant-Seeker που έχει τη δυνατότητα να εντοπίζει συναφείς πληροφοριακές μονάδες δρομολογώντας την αναζήτηση πληροφορίας μέσα στο δυναμικό περιβάλλον του Παγκόσμιου Ιστού. Η δρομολόγηση της αναζήτησης, πραγματοποιείται στοχαστικά συνδυάζοντας τεχνικές ανάκτησης που βασίζονται στην ομοιότητα εγγράφων και τεχνικών προσομοίωσης του τρόπου επικοινωνίας των μυρμηγκιών. Αφού παρουσιαστούν αρχικά οι βασικότερες εμπνευσμένες από τη φύση τεχνικές, αναλύοντας τους περιορισμούς των παραδοσιακών μεθόδων μηχανικής μάθησης καθώς και τους λόγους που οδήγησαν στη αναζήτηση τέτοιων εναλλακτικών, μη συμβατικών μεθόδων, παρουσιάζονται οι αρχές δυο θεμελιωδών μεθόδων εμπνευσμένων από τη φύση. Στη συνέχεια παρουσιάζεται ο προτεινόμενος αλγόριθμος και το κεφάλαιο κλείνει με την παρουσίαση πειραματικών μετρήσεων για την προτεινόμενη μεθοδολογία καθώς και ποιοτικά συμπεράσματα.

4.2 Τεχνικές Νοημοσύνης Σμήνους

Ο όρος Νοημοσύνη Σμήνους (Swarm Intelligence) εμφανίστηκε για πρώτη φορά στην εργασία [109] στα πλαίσια ερευνών σχετικές με τη ρομποτική. Γενικά, οι μέθοδοι και οι αλγόριθμοι που υπάγονται σε αυτή την κατηγορία, αντλούν έμπνευση από τη συμπεριφορά εντόμων, πτηνών και ψαριών (ή γενικότερα σχηματισμών που προσομοιάζουν σμήνη από ζώα). Τέτοιοι σχηματισμοί παρουσιάζουν μοναδική ικανότητα στην περάτωση πολύπλοκων εργασιών όταν δρουν συλλογικά, παρόλο που κάτι τέτοιο θα ήταν αδύνατο αν δρούσαν σε ατομικό επίπεδο. Πράγματι, ένα μυρμήγκι,

μέλισσα, πτηνό ή ψάρι διαθέτει πολύ περιορισμένη νοημοσύνη ατομικά αλλά όταν συναναστρέφεται με άλλες μονάδες του είδους του, καταφέρνει να ολοκληρώσει πιο περίπλοκες εργασίες όπως το να βρει το γρηγορότερο μονοπάτι προς μια πηγή τροφής, να οργανώσει με βέλτιστο τρόπο την φωλιά του, να συγχρονίσει τις κινήσεις με όμοιούς του και να κινηθεί ταχύτατα σε σχηματισμούς. Τα κατορθώματα αυτά είναι ακόμα πιο θαυμαστά αν αναλογιστεί κανείς ότι αυτά επιτυγχάνονται χωρίς την άμεση επικοινωνία μέσω λόγου και χωρίς κάποια κεντρική αρχή (π.χ. η βασίλισσα της φωλιάς) να επιβλέπει και να οργανώνει τις κινήσεις του σμήνους/κοινωνίας. Μοντελοποιήσεις αυτής της συμπεριφοράς έχουν βρει εφαρμογή σε προβλήματα όπως ο χρονοπρογραμματισμός διεργασιών, η δρομολόγηση οχημάτων κ.α.

Τα μοναδικά χαρακτηριστικά των τεχνικών Νοημοσύνης Σμήνους, έχουν αναδείξει αυτή την οικογένεια αλγορίθμων ως μια από τις καλύτερες επιλογές μεταξύ των υπαρχόντων για μηχανική μάθηση. Τα πλεονεκτήματα μπορούν να γίνουν εύκολα αντιληπτά αν αναλογιστεί κανείς την αντιστοιχία με τα φυσικά συστήματα: οι αλγόριθμοι Νοημοσύνης Σμήνους στοχεύουν στην επίλυση δύσκολων προβλημάτων βασιζόμενοι σε πολλαπλούς αλλά απλούς δομικά, πράκτορες, οι οποίοι δεν απαιτούν κεντρική αρχή για την οργάνωσή τους. Τέτοιοι πράκτορες συνεργάζονται για την ανεύρεση μιας βέλτιστης λύσης για ένα συγκεκριμένο πρόβλημα. Η οργάνωσή τους είναι φυσική απόρροια της έμμεσης επικοινωνίας (η οποία συνήθως επιτυγχάνεται σημαδεύοντας το περιβάλλον μέσα στο οποίο δρουν). Έτσι, τέτοιοι πράκτορες μπορούν να χρησιμοποιηθούν για ανακάλυψη ποιοτικών κανόνων κατηγοριοποίησης ή για τη δημιουργία ομογενών συστάδων. Τέτοια συμπεριφορά είναι ιδιαίτερα επιθυμητή καθώς μπορεί να οδηγήσει σε συστήματα που λειτουργούν παράλληλα, και είναι προσαρμόσιμα σε πιθανές αλλαγές συνθηκών ενώ είναι και αποδοτικά από άποψη κόστους. Δυο τεχνικές που εντάσσονται στην ευρύτερη οικογένεια των τεχνικών Νοημοσύνης Σμήνους είναι: α) η Βελτιστοποίηση Αποικίας Μυρμηγκιών και β) η Βελτιστοποίηση Σμήνους Στοιχείων.

4.2.1 Βελτιστοποίηση Αποικίας Μυρμηγκιών

Η συμπεριφορά ορισμένων ειδών μυρμηγκιών κατά την αναζήτηση της τροφής τους και πιο συγκεκριμένα η μοναδική ικανότητα που έχουν να ανακαλύπτουν το συντομότερο μονοπάτι προς μια πηγή τροφής, αποτέλεσε πηγή έμπνευσης για τον πιο δημοφιλή αλγόριθμο Νοημοσύνης Σμήνους, τον επονομαζόμενο αλγόριθμο Βελτιστοποίησης Αποικίας Μυρμηγκιών (Ant Colony Optimization).

Τα περισσότερα είδη μυρμηγκιών έχουν πολύ περιορισμένη ή ακόμα και καθόλου ορατότητα, ενώ ταυτόχρονα δεν διαθέτουν ομιλία ή άλλα μέσα άμεσης επικοινωνίας. Παρ' όλα αυτά οι κινήσεις των μυρμηγκιών μοιάζουν να είναι απόλυτα οργανωμένες, γεγονός που δηλώνει ότι λαμβάνει χώρα κάποιου είδους επικοινωνία λιγότερο εμφανής. Πράγματι, σχετικά πειράματα τα οποία διεξήχθησαν σε ορισμένα είδη αποδεικνύει ότι αυτή η επικοινωνία πραγματοποιείται με τη βοήθεια μιας ειδικής ουσίας που ονομάζεται φερομόνη, την οποία τα μυρμήγκια εναποθέτουν στη διαδρομή τους.

Αναλυτικότερα η διαδικασία είναι η ακόλουθη: Αρχικά τα μυρμήγκια περιφέρονται κατά τρόπο τυχαίο αναζητώντας μια πηγή τροφής. Όταν αυτό συμβεί τα μυρμήγκια μεταφέρουν την τροφή πίσω στη φωλιά τους εναποθέτοντας παράλληλα και μια ποσότητα φερομόνης κατά μήκος αυτής της διαδρομής. Από το σημείο αυτό και μετά τα μυρμήγκια αποφασίζουν ποια από τις υπάρχουσες διαδρομές θα ακολουθήσουν με βάση τη συγκέντρωση της φερομόνης στην εκάστοτε διαδρομή. Όπως είναι φυσικό οι διαδρομές με μεγαλύτερη συγκέντρωση φερομόνης έχουν μεγαλύτερη πιθανότητα να επιλεγούν. Τα μυρμήγκια που ακολουθούν τη συντομότερη διαδρομή επιστρέφουν στη φωλιά τους πιο σύντομα και έτσι η φερομόνη σε αυτό το μονοπάτι ενισχύεται με επιπλέον ποσότητες πιο γρήγορα σε σύγκριση με τα μεγαλύτερα μονοπάτια. Κατ' αυτό τον τρόπο η επιλογή μεταξύ των υποψήφιων μονοπατιών κλίνει πάντα προς το συντομότερο μονοπάτι.

Στην εργασία [110] παρουσιάζεται το πείραμα της διπλής γέφυρας στο οποίο μια αποικία από μυρμήγκια καλούνταν να περάσουν μια γέφυρα από δυο μονοπάτια ίσου μήκους. Στα πειράματα αυτά παρατηρήθηκε ότι ενώ αρχικά ο πληθυσμός επέλεγε ένα μονοπάτι τυχαία, με την πάροδο του χρόνου όλα τα μυρμήγκια της αποικίας κατέληγαν στο να ακολουθούν μόνο ένα εκ των δυο. Ποιο από τα δυο, καθορίζεται τυχαία. Στην εργασία [111] το πείραμα της διπλής γέφυρας επεκτάθηκε κάνοντας χρήση μονοπατιών διαφορετικού μήκους και παρατηρήθηκε ότι σε όλες τις περιπτώσεις, τελικά, τα μυρμήγκια πάντα επιλέγουν το συντομότερο μονοπάτι όπως φαίνεται στην Εικόνα 4.1.

Εικόνα 4.1 - Στο πείραμα διπλής γέφυρας τα μυρμήγκια επιλέγουν το συντομότερο μονοπάτι μετά από κάποιο χρονικό διάστημα

Στην εργασία [112] παρουσιάζεται ένα αλγοριθμικό μοντέλο αυτής της συμπεριφοράς των μυρμηγκιών με την ονομασία Απλή Βελτιστοποίηση Αποικίας Μυρμηγκιών (Simple Ant Colony Optimization) που έχει σαν στόχο την εφαρμογή του σε προβλήματα ελαχιστοποίησης κόστους διαδρομής σε γράφους. Σε αυτό το μοντέλο τα μυρμήγκια ξεκινούν από ένα αρχικό κόμβο του γράφου $G = (N, A)$ και επιχειρούν να φτάσουν σε ένα τελικό κόμβο ακολουθώντας το πιο σύντομο μονοπάτι. Σε κάθε ακμή (i, j) του γράφου εναποτίθεται ένα ποσό συμβολικής φερομόνης $\tau_{i,j}$. Αυτή η πληροφορία μπορεί να διαβαστεί και να μεταβληθεί από τα μυρμήγκια ώστε να επηρεάσει την επιλογή του επόμενου κόμβου προς επίσκεψη.

Πιο συγκεκριμένα, η πιθανότητα που έχει ένα μυρμήγκι k το οποίο βρίσκεται σε ένα κόμβο i να επιλέξει τον κόμβο j ως τον επόμενο κόμβο προς επίσκεψη υπολογίζεται σύμφωνα με την εξίσωση (4.1):

$$p_{ij}^{k} = f(x) = \begin{cases} \dfrac{\tau_{ij}^{a}}{\sum_j \tau_{ij}^{a}}, & j \in N_i^k \\ 0, & j \notin N_i^k \end{cases} \tag{4.1}$$

Όπου N_i^k είναι όλοι οι κόμβοι που συνδέονται άμεσα με το i εκτός από τον προηγούμενο και α είναι μια παράμετρος που ελέγχει την ταχύτητα σύγκλησης. Όταν το μυρμήγκι φτάσει στον προορισμό του θα πρέπει να επιστρέψει πίσω στον αρχικό κόμβο. Σε αυτή την αντίστροφη πορεία τα μυρμήγκια εναποθέτουν φερομόνη. Τυπικά κάθε μυρμήγκι θα προσπαθήσει να ακολουθήσει την ίδια πορεία όμως αυτή μπορεί να περιέχει βρόγχους οπότε και αυτοί θα πρέπει να απαλειφθούν ώστε να μην ενισχυθούν από την φερομόνη. Η νέα συγκέντρωση φερομόνης σε μια ακμή (i, j) όταν ένα μυρμήγκι k την επισκέφτηκε ως μέρος της διαδρομής επιστροφής προς τον αρχικό κόμβο υπολογίζεται ως εξής:

$$\tau_{ij} \leftarrow \tau_{ij} + \Delta\tau^{k} \tag{4.2}$$

Παράλληλα με τη συγκέντρωση φερομόνης, στους κόμβους των διαδρομών που δεν επισκέφτηκαν μυρμήγκια συμβαίνει και η εξάτμισή της. Ο μηχανισμός εξάτμισης της φερομόνης μπορεί να θεωρηθεί ως ένας τρόπος για να αποφευχθεί η σύγκληση σε μη βέλτιστα μονοπάτια, ή σαν ένας τρόπος προσαρμογής σε δυναμικές αλλαγές στο γράφο αν προκύψουν ποτέ τέτοιες. Η εξάτμιση της φερομόνης προσομοιάζεται εφαρμόζοντας την ακόλουθη εξίσωση σε όλες τις ακμές:

$$\tau_{ij} \longleftarrow (1-p)\tau_{ij} \forall (i,j) \in A \tag{4.3}$$

Όπου $p \in (0,1]$ είναι μια σταθερά.

4.2.2 Βελτιστοποίηση Σμήνους Σωματιδίων

Η τεχνική Βελτιστοποίησης Σμήνους Στοιχείων (Particle Swarm Optimization) αντλεί έμπνευση από τη συγχρονισμένη κίνηση που παρατηρείται σε σμήνη από ζώα όπως είναι τα πτηνά ή τα ψάρια. Βάσει μελετών που διεξήχθησαν στην εργασία [113] η κίνηση ενός σμήνους είναι αποτέλεσμα των ατομικών ενεργειών των πτηνών. Οι ενέργειες αυτές υπακούν σε τρείς βασικούς κανόνες:

- Την αποφυγή συγκρούσεων, που επιτάσσει κάθε στοιχείο του σμήνους να αναπροσαρμόζει τη θέση του ώστε να αποφύγει τη σύγκρουση με ένα άλλο γειτονικό πτηνό.
- Την ταύτιση ταχύτητας, που επιβάλει σε κάθε πτηνό του σμήνους να συγχρονίζει την ταχύτητά του με αυτή των γειτόνων του.
- Το κεντράρισμα, που επιβάλει στα πτηνά να παραμένουν κοντά και σε σχετικά σταθερή θέση από τους γείτονές τους.

Το παραπάνω απλό μοντέλο εφαρμόστηκε για να εξομοιώσει σε ένα τρισδιάστατο γραφικό περιβάλλον τη χωρογραφία ενός σμήνους από πτηνά στην οθόνη του υπολογιστή. Σε μια παλιότερη

μελέτη [114] είχε διατυπωθεί η εξής παρατήρηση: τα μέλη ενός σμήνους πουλιών μπορεί να επωφεληθούν από τις ανακαλύψεις και τις προηγούμενες εμπειρίες ξεχωριστών μελών του σμήνους. Με άλλα λόγια, ένα μεγαλύτερο σμήνος αυξάνει τις πιθανότητες ανεύρεσης μιας πηγής πλούσιας σε τροφή και η ανταλλαγή πληροφοριών μεταξύ των μελών ενός σμήνους προσφέρει μεγαλύτερο πλεονέκτημα. Ωστόσο, η εισαγωγή του όρου «Βελτιστοποίηση Σμήνους Σωματιδίου» και του αντίστοιχου αλγοριθμικού μοντέλου έγινε στην εργασία [115].

Το συγκεκριμένο μοντέλο υποθέτει την ύπαρξη μιας συνάρτησης καταλληλότητας $f: R^n \rightarrow R$ η οποία αξιολογεί την ποιότητα μιας λύσης. Ένας αριθμός S από σωματίδια τοποθετούνται σε ένα υπερχώρο σε τυχαία θέση $x \in R^n$ έχοντας το καθένα τυχαία ταχύτητα $u_i \in R^n$ Τα σωματίδια κινούνται στον υπερχώρο και σε κάθε βήμα αξιολογούν τη θέση τους με την συνάρτηση καταλληλότητας. Κάθε σωματίδιο στο σμήνος αναπαριστά μια πιθανή λύση και η κίνησή τους ισοδυναμεί με ελαφρά τροποποίηση της λύσης. Ο βασικός κανόνας για την αλλαγή της ταχύτητας είναι:

$$u_i(t+1) = \omega u_i(t) + c_1 r_1 (p_i - x_i) + c_2 r_2 (g - x_i) \tag{4.4}$$

Όπου ω είναι μια σταθερά που αναπαριστά το βάρος αδράνειας, c_1, c_2 είναι σταθερές που αναπαριστούν την επιτάχυνση, r_1, r_2 είναι τυχαίοι αριθμοί, p_i είναι η καλύτερη θέση του σωματιδίου i, g είναι η καλύτερη θέση μεταξύ όλων των σωματιδίων στο σμήνος, και x_i είναι η τρέχουσα θέση του σωματιδίου i.

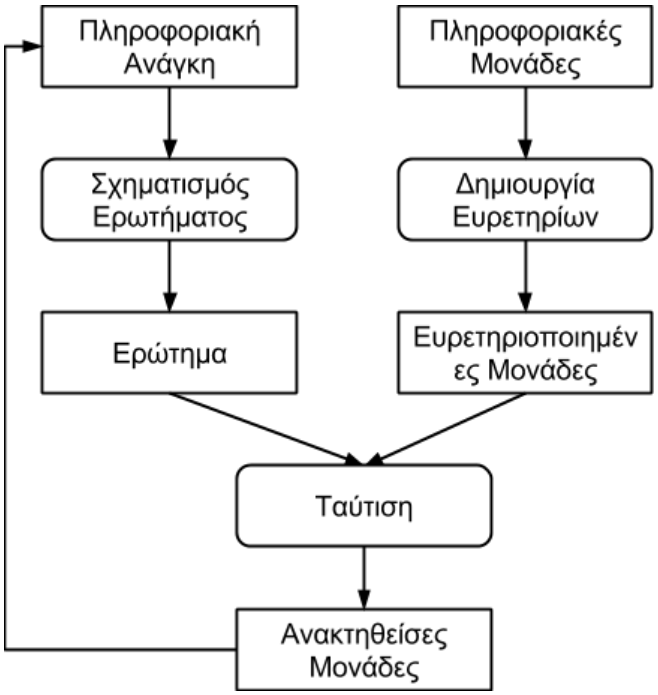
Επιπλέον, ο κανόνας για την αλλαγή της θέσης είναι:

$$x_i(t+1) = x_i + u_i(t+1) \tag{4.5}$$

Τα δύο βασικά χαρακτηριστικά του μοντέλου είναι το γεγονός ότι: (α) η ταχύτητα (και επομένως η επόμενη θέση) των σωματιδίων καθορίζεται από τα ευρήματα τόσο του ίδιου του σωματιδίου όσο και από τα ευρήματα ολόκληρου του σμήνους, και (β) η καθολική βέλτιστη θέση των σωματιδίων του σμήνους γνωστοποιείται σε όλα τα σωματίδια του σμήνους. Η μέθοδος Βελτιστοποίησης Σμήνους Σωματιδίων παρουσιάζει αρκετές ομοιότητες με τη μέθοδο Γενετικών Αλγορίθμων. Πράγματι, και οι δυο μέθοδοι λαμβάνουν υπόψη μια συνάρτηση καταλληλότητας η οποία δρα σαν κριτήριο για την αναπαραγωγή του πληθυσμού και μεταβάλουν τον πληθυσμό με τυχαίο τρόπο. Ωστόσο, η μέθοδος Βελτιστοποίησης Σμήνους Σωματιδίων δεν στηρίζεται σε γενετικές μεθόδους αλλαγής όπως είναι η μετάλλαξη και η γονιδιακή αλλαγή. Επιπλέον, η μέθοδος Βελτιστοποίησης Σμήνους Σωματιδίων απαιτεί ένα είδος μηχανισμού μνήμης ως απαραίτητο στοιχείο για τη σύγκληση σε μια βέλτιστη λύση.

4.3 Ανάκτηση πληροφορίας και τεχνικές αναπαράστασης εγγράφων

Η Ανάκτηση Πληροφορίας (Information Retrieval) ασχολείται με την αναπαράσταση, την αποθήκευση, την οργάνωση πληροφοριακών μονάδων και την πρόσβαση σε αυτές. Απώτερος στόχος της Ανάκτησης Πληροφορίας είναι η ικανοποίηση των πληροφοριακών αναγκών του ανθρώπινου παράγοντα. Όπως φαίνεται στην Εικόνα 4.2, η διαδικασία της ανάκτησης μιας πληροφορίας ξεκινά με την υποβολή ενός ερωτήματος από το χρήστη στο σύστημα ανάκτησης. Τα ερωτήματα είναι εκφράσεις των πληροφοριακών αναγκών των χρηστών και μπορεί να έχουν τη μορφή φράσεων συζευγμένων όρων. Ένα ερώτημα μπορεί να μην προσδιορίζει μοναδικά μια πληροφοριακή μονάδα σε μια συλλογή, άρα μπορεί να συσχετίζεται με πολλαπλές μονάδες, ενδεχομένως μάλιστα και με διαφορετικό βαθμό σχετικότητας.



Εικόνα 4.2 - Η διαδικασία της ανάκτησης πληροφορίας

Μια πληροφοριακή μονάδα είναι μια οντότητα που αναπαρίσταται από δεδομένα αποθηκευμένα σε μια βάση δεδομένων τα οποία συγκρίνονται με το ερώτημα κατά την υποβολή του. Ανάλογα με την εφαρμογή, οι μονάδες μπορεί να είναι έγγραφα, εικόνες ή άλλα πολυμεσικά δεδομένα. Συχνά οι μονάδες αυτές δεν αποθηκεύονται στην αρχική μορφή τους, αλλά σε μια πιο περιληπτική μορφή, ή ως μεταδεδομένα. Τα περισσότερα συστήματα ανάκτησης πληροφορίας υπολογίζουν τον βαθμό ταύτισης κάθε μονάδας με το ερώτημα που υποβάλλεται από το χρήστη και κατατάσσουν τις μονάδες ανάλογα με το βαθμό αυτό. Οι υψηλότερες μονάδες στην κατάταξη εμφανίζονται στο χρήστη.

4.3.1 Ανάκτηση πληροφορίας στον Παγκόσμιο Ιστό

Ο αρχικός προορισμός των συστημάτων ανάκτησης πληροφορίας αφορούσε την αυτοματοποιημένη ευρετηριοποίηση συναφών πληροφοριακών μονάδων και την ανάπτυξη μεθόδων για την αναζήτησή τους σε μια συλλογή. Η κυριότερη εφαρμογή που έβρισκε μέχρι τις αρχές τις δεκαετίας του 90 ήταν στην περιοχή της βιβλιοθηκονομίας. Ωστόσο αυτό άλλαξε δραματικά με την εμφάνιση του Παγκόσμιου Ιστού, που ανέδειξε νέα ζητήματα όπως τη μοντελοποίηση, την κατηγοριοποίηση και την οπτικοποίηση των δεδομένων αλλά και της αρχιτεκτονικής των συστημάτων ανάκτησης πληροφορίας για να είναι σε θέση να ανταποκριθούν στην τεράστια κλίμακα του Παγκόσμιου Ιστού.

Ο Παγκόσμιος Ιστός αποτελεί τη μεγαλύτερη πηγή πληροφοριών και τη μεγαλύτερη συλλογή ανθρώπινης γνώσης. Η επιτυχία του έγκειται στην ευκολία που παρέχει στον τελικό χρήστη για τη δημιουργία περιεχομένου, καθιστώντας τον ένα εύκολα προσπελάσιμο και φτηνό μέσο έκφρασης. Ο Παγκόσμιος Ιστός επιπρόσθετα, θέτει εναλλακτικούς τρόπους επικοινωνίας που καταργούν πολλές φορές τις έννοιες της χωροχρονικής απόστασης. Τέλος, δραστηριότητες όπως το ηλεκτρονικό εμπόριο και η ενημέρωση, έχουν πλέον εξελιχθεί σε τέτοιο βαθμό, που αποτελούν αναπόσπαστο κομμάτι της καθημερινότητας εκατομμυρίων χρηστών.

Από τη σκοπιά του χρήστη, η εύρεση χρήσιμης πληροφορίας στον Παγκόσμιο Ιστό είναι μια σχετικά δύσκολη και χρονοβόρος διαδικασία. Μια τυχαία περιπλάνηση του χρήστη από ιστοσελίδα σε ιστοσελίδα, ακολουθώντας τους υπερσυνδέσμους που βρίσκονται μέσα τους, σε μια προσπάθεια να εντοπίσει την επιθυμητή πληροφορία είναι συχνά αναποτελεσματική, αφού το μέγεθος του Παγκόσμιου Ιστού είναι απαγορευτικό, ενώ ο χρήστης συχνά δεν γνωρίζει ένα καλό σημείο εκκίνησης. Για τους άπειρους χρήστες, το πρόβλημα της αναζήτησης γίνεται πολύ πιο δύσκολο και συχνά τους οδηγεί σε απογοητευτικά αποτελέσματα. Το κύριο εμπόδιο, είναι η απουσία ενός καλά ορισμένου μοντέλου δεδομένων για τον Παγκόσμιο Ιστό, κάτι που έχει σαν αποτέλεσμα τον ορισμό και τη δόμηση της πληροφορίας να είναι χαμηλής ποιότητας. Αυτές οι δυσκολίες έστρεψαν το ενδιαφέρον στον τομέα της Ανάκτησης Πληροφορίας και οδήγησαν στην υιοθέτηση δοκιμασμένων τεχνικών που χρησιμοποιούνται εκεί, ως πολλά υποσχόμενων λύσεων.

4.3.2 Μοντέλα ανάκτησης πληροφορίας

Η ύπαρξη των μοντέλων ανάκτησης πληροφορίας οφείλεται αφενός στο ότι παρέχουν ένα σημείο αναφοράς για μελέτη και την καθοδήγηση της έρευνας για την εξέλιξή τους και αφετέρου γιατί αποτελούν προσχέδιο για τη δημιουργία πραγματικών συστημάτων ανάκτησης πληροφορίας.

Τα μαθηματικά μοντέλα χρησιμοποιούνται σε πολλές ερευνητικές περιοχές με στόχο την κατανόηση και το συλλογισμό πάνω σε μια παρατήρηση, συμπεριφορά ή φαινόμενο του πραγματικού κόσμου. Ένα μοντέλο στην Ανάκτηση Πληροφορίας ωστόσο, προβλέπει και αιτιολογεί

τι θα βρει χρήσιμο ένας χρήστης δεδομένων των πληροφοριακών αναγκών του. Η ορθότητα των προβλέψεων του μοντέλου μπορούν να δοκιμαστούν σε ένα ελεγχόμενο περιβάλλον. Τα μοντέλα Ανάκτησης Πληροφορίας βασίζονται σε σκεπτικά, μεταφορές και μαθηματικά υπόβαθρα, για να είναι δυνατή η καλύτερη κατανόησή τους. Με το σκεπτικό ενός μοντέλου καθίσταται δυνατή η αποδοχή ή η απόρριψή τους από την ερευνητική κοινότητα. Με τις μεταφορές γίνεται πιο εύκολη η περιγραφή τους σε ένα ευρύτερο κοινό, ενώ με τους μαθηματικούς φορμαλισμούς εξασφαλίζεται η συνέπεια και η διασφάλιση της δυνατότητας υλοποίησής τους.

Ένας πιο επίσημος ορισμός για την έννοια του μοντέλου Ανάκτησης Πληροφορίας δίνεται στο [116]. Ένα μοντέλο λοιπόν είναι η τετράδα [D, Q, F, $R(q_i,d_j)$], όπου το D είναι ένα σύνολο από λογικές αναπαραστάσεις για τις πληροφοριακές μονάδες της συλλογής, το Q αντιπροσωπεύει ένα σύνολο από λογικές αναπαραστάσεις για τις πληροφοριακές ανάγκες (ερωτήσεις) του χρήστη, το F αποτελεί το υπόβαθρο για την μοντελοποίηση της αναπαράστασης των πληροφοριακών μονάδων, των ερωτημάτων και των σχέσεων μεταξύ τους και το $R(q_i,d_j)$, είναι μια συνάρτηση που συνδέει έναν πραγματικό αριθμό με ένα ερώτημα $q_i \in Q$ και μια αναπαράσταση κειμένου $d_j \in D$. Μια τέτοια κατάταξη ορίζει μια διάταξη πάνω στις πληροφοριακές μονάδες πάντα βάσει του ερωτήματος q_i.

Τα κυριότερα μοντέλα στην Ανάκτηση Πληροφορίας είναι τρία:

- το συνολοθεωρητικό (set-theoretic) όπου τόσο οι πληροφοριακές μονάδες, όσο και τα ερωτήματα αντιμετωπίζονται σαν σύνολα από όρους. Τα κυριότερα μοντέλα που ανήκουν σε αυτή την κατηγορία είναι το Κλασικό και το Εκτεταμένο Boolean [117] και το Ασαφές (Fuzzy) [118].
- το αλγεβρικό (algebraic) όπου οι μονάδες και τα ερωτήματα αναπαρίστανται ως διανύσματα, πίνακες ή πλειάδες ενός πολυδιάστατου χώρου. Τα κυριότερα μοντέλα αυτής της κατηγορίας είναι το Χωροδιανυσματικό Μοντέλο (Vector space model) και η Λανθάνουσα Σημασιολογική Ευρετηριοποίηση (Latent Semantic Indexing) [119] .
- το πιθανοτικό (probabilistic) όπου η ανάκτηση εκφράζεται μέσω μιας πιθανοτικής επαγωγής. Οι ομοιότητες υπολογίζονται βάσει των πιθανοτήτων των μονάδων να είναι σχετικές με δεδομένα ερωτήματα. Κυριότερα μοντέλα αυτής της κατηγορίας είναι το μοντέλο Δυαδικής Ανεξαρτησίας (Binary Independence) [120] και τα γλωσσικά μοντέλα [121].

Από τα παραπάνω μοντέλα αναλύεται στην επόμενη παράγραφο το Χώρο-διανυσματικό, αφού σε αυτό βασίζεται η προτεινόμενη προσέγγιση του κεφαλαίου.

4.3.2.1 Το Χώρο-Διανυσματικό μοντέλο

Το Χώρο-Διανυσματικό μοντέλο (Vector Space Model) είναι ένα αλγεβρικό μοντέλο ανάκτησης πληροφορίας για την αναπαράσταση εγγράφων (αλλά και άλλων αντικειμένων γενικότερα) με τη

μορφή διανυσμάτων όρων. Πέραν της ανάκτησης πληροφορίας, χρησιμοποιείται στην ευρετηριοποίηση και την διάταξη ομοιοτήτων. Τα έγγραφα και τα ερωτήματα σε αυτό το μοντέλο αναπαρίστανται ως διανύσματα με διάσταση t: $d_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$ και $q = (w_{1,q}, w_{2,q}, \dots, w_{t,q})$.

Στο μοντέλο αυτό ο βαθμός ομοιότητας μεταξύ του αντικειμένου d_j και του ερωτήματος q υπολογίζεται ως ο βαθμός συσχέτισης μεταξύ των δυο διανυσμάτων. Μέτρο του βαθμού συσχέτισης αποτελεί το συνημίτονο της γωνίας που σχηματίζεται από τα δύο διανύσματα και παρέχεται από τη σχέση (4.6):

$$sim(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \|q\|} \tag{4.6}$$

Όπου $d_j \cdot q$ είναι το εσωτερικό γινόμενο του διανύσματος του εγγράφου και του διανύσματος του ερωτήματος και $\|d_j\|, \|q\|$ οι νόρμες των διανυσμάτων αντίστοιχα. Η νόρμα ενός διανύσματος υπολογίζεται ως: $\|q\| = \sqrt{\sum_{i=1}^{n} q_i}$. Εφόσον όλα τα διανύσματα που λαμβάνονται υπόψη σε αυτό το μοντέλο είναι θετικά, αν το συνημίτονο ισούται με μηδέν τότε το ερώτημα και το έγγραφο είναι ορθογώνια και δεν ταυτίζονται καθόλου (δηλαδή ο όρος του ερωτήματος δεν υπάρχει στο έγγραφο που λαμβάνεται υπόψη), ενώ αντίθετα αν ισούται με τη μονάδα το ερώτημα και το έγγραφο ταυτίζονται πλήρως.

Στο κλασικό χώρο-διανυσματικό μοντέλο, τα βάρη των διανυσμάτων των εγγράφων είναι γινόμενα τοπικών και καθολικών παραμέτρων. Οι παράμετροι αυτοί είναι α) η συχνότητα των όρων (term frequency – TF), που αποτυπώνει τη διαίσθηση ότι όσο πιο συχνά εμφανίζεται ένας όρος σε ένα κείμενο, τόσο πιο καλή περιγραφή του κειμένου θα αποτελεί και β) η αντίστροφη συχνότητα εμφάνισης (inverse document frequency – IDF) που αποτυπώνει τη διαίσθηση ότι όσο μεγαλύτερη συχνότητα εμφάνισης έχει ένας όρος σε μια συλλογή, τόσο λιγότερο χρήσιμος είναι για να χαρακτηρίσει ένα κείμενο και άρα να το διαχωρίσει μέσα σε μια συλλογή.

Έστω N ο συνολικός αριθμός των κειμένων και n_i ο αριθμός των κειμένων στα οποία εμφανίζεται ο όρος k_i. Έστω f_{ij} η συχνότητα εμφάνισης του όρου k_i στο d_j. Τότε η κανονικοποιημένη συχνότητα TF_{ij} του όρου k_i στο d_j δίνεται από τη σχέση (4.7), όπου η μέγιστη τιμή max υπολογίζεται πάνω σε κάθε όρο που αναφέρεται στο κείμενο d_j. Αν ο όρος k_i δεν εμφανίζεται στο d_j τότε $TF_{ij}=0$.

$$TF_{i,j} = \frac{f_{i,j}}{max f_{i,j}} \tag{4.7}$$

Επιπλέον, έστω IDF_i η αντίστροφη συχνότητα εμφάνισης για τον όρο k_i που δίνεται από τη σχέση (4.8).

$$IDF_i = \log \frac{N}{n_i} \tag{4.8}$$

Ο συνδυασμός των δυο αυτών μεγεθών ορίζει το μοντέλο ανάθεσης βαρών TF-IDF, σύμφωνα με τη σχέση (4.9). Αντίστοιχα, για τα βάρη των όρων στα ερωτήματα ισχύει η σχέση (4.10) που είναι μια διπλή κανονικοποίηση 0.5, για να αποφεύγεται η προτίμηση σε μεγαλύτερα έγγραφα.

$$w_{i,j} = TF_{i,j} IDF_i = \frac{f_{i,j}}{max f_{i,j}} \log \frac{N}{n_i} \tag{4.9}$$

$$w_{i,q} = \left(0.5 + 0.5 \frac{f_{i,j}}{max f_{i,j}}\right) \log \frac{N}{n_i} \tag{4.10}$$

Αντικαθιστώντας τη σχέση (4.6) με τα παραπάνω βάρη τελικά η ομοιότητα ερωτήματος-εγγράφου ορίζεται ως:

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} (w_{i,j} w_{i,q})}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \tag{4.11}$$

Με το μοντέλο αυτό είναι δυνατή η κατάταξη των εγγράφων με κριτήριο το βαθμό ομοιότητάς τους προς το ερώτημα. Επίσης μπορεί να οριστεί μια ελάχιστη τιμή ομοιότητας, σύμφωνα με την οποία θα επιστρέφονται τα έγγραφα στο χρήστη ως σχετικά ή θα αποκλείονται.

Το χώρο-διανυσματικό μοντέλο παρουσιάζει μια σειρά πλεονεκτημάτων ορισμένα από τα οποία είναι: α) είναι ένα απλό μοντέλο βασισμένο στη γραμμική άλγεβρα, β) τα βάρη των όρων δεν είναι δυαδικά, γ) επιτρέπει τον υπολογισμό ενός συνεχούς βαθμού ομοιότητας μεταξύ των ερωτημάτων και των εγγράφων, δ) επιτρέπει την κατάταξη των αποτελεσμάτων ανάλογα με τη σχετικότητά τους με το ερώτημα και ε) επιτρέπει τη μερική ταύτιση.

Ωστόσο το χώρο-διανυσματικό μοντέλο παρουσιάζει και μια σειρά μειονεκτημάτων: α) τα μεγάλα έγγραφα δεν αντιπροσωπεύονται επαρκώς επειδή έχουν χαμηλές τιμές ομοιότητας (μικρό βαθμωτό γινόμενο και μεγάλες διαστάσεις) β) οι όροι πρέπει να ταυτίζονται ακριβώς, κάτι που σημαίνει ότι ακόμα και ένα γράμμα διαφορετικό μεταξύ των όρων οδηγεί σε μη-ταύτιση, γ) η σειρά με την οποία εμφανίζονται οι όροι στο κείμενο χάνεται στην αναπαράσταση του διανυσματοχώρου, δ) υποθέτει θεωρητικά ότι οι όροι είναι στατιστικά ανεξάρτητοι και ε) είναι σημασιολογικά ευαίσθητο, κάτι που σημαίνει ότι τα έγγραφα με παρόμοιο περιεχόμενο, αλλά διαφορετικούς όρους δεν θα συσχετιστούν και στ) η ανάθεση των βαρών δεν είναι επίσημη και για το λόγο αυτό υπάρχουν διάφορες παραλλαγές του μοντέλου.

4.4 Η προτεινόμενη προσέγγιση

Η προτεινόμενη προσέγγιση, γνωστή ως Ant-Seeker, στηρίζεται στη θεωρητική αρχή σύγκλισης και μια τροποποίηση του βασικού μοντέλου αποικίας μυρμηγκιών που περιγράφτηκε στην

παράγραφο 4.2.1. Η διαφοροποίηση του προτεινόμενου αλγορίθμου είναι σε σύγκριση με τους περισσότερους συμβατικούς αλγορίθμους της οικογένειας Αποικίας Μυρμηγκιών, έγκειται στο γεγονός ότι μπορεί επιτυχώς να εκτελέσει αναζήτηση σε ένα δυναμικό περιβάλλον, η δομή του οποίου δεν είναι καθορισμένη εξαρχής, όπως είναι ο Παγκόσμιος Ιστός.

Η προτεινόμενη προσέγγιση αναπαριστά τις ιστοσελίδες ενός τμήματος του Παγκόσμιου Ιστού ως κόμβους ενός γράφου, του οποίου οι ακμές είναι οι υπερσύνδεσμοι (hyperlinks) που ενώνουν τις ιστοσελίδες αυτές.. Ο Παγκόσμιος Ιστός λοιπόν μοντελοποιείται σαν ένας γράφος απείρων διαστάσεων $G_p = (P, L)$ όπου P οι κόμβοι του γράφου που αναπαριστούν τις ιστοσελίδες και L οι ακμές του γράφου που αναπαριστούν τους υπερσυνδέσμους μεταξύ των σελίδων αυτών. Σε αντίθεση με το μεγαλύτερο όγκο της υπάρχουσας βιβλιογραφίας, που προτείνουν λύσεις στο πρόβλημα ανεύρεσης της πιο σύντομης διαδρομής μεταξύ δυο κόμβων, το συγκεκριμένο πρόβλημα ανάγεται στην αναζήτηση όμοιων κόμβων στο γράφο.

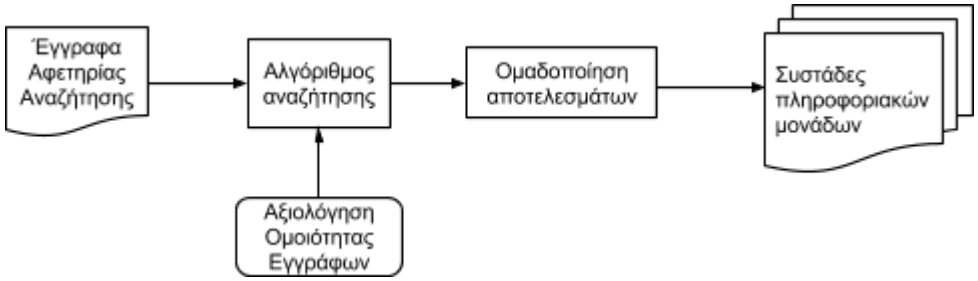
Στην προτεινόμενη προσέγγιση, γίνεται η παραδοχή ότι αν υπάρχει ένας κόμβος Α στον γράφο ο οποίος περιέχει μια ιστοσελίδα σχετική με την πληροφοριακή ανάγκη του χρήστη, τότε θα υπάρχει ένας άλλος κόμβος Β σε "κοντινή απόσταση", ο οποίος θα περιέχει σχετική με αυτόν πληροφορία [122]. Στην προκειμένη περίπτωση, ως μονάδα απόστασης ορίζεται ο αριθμός των ακμών που χωρίζει τους δύο κόμβους. Αυτός, αντικατοπτρίζει τον αριθμό των υπερσυνδέσμων που απαιτείται να ακολουθηθεί ώστε να γίνει η μετάβαση από την ιστοσελίδα Α σε μια ιστοσελίδα Β.

Από μια προοπτική υψηλού επιπέδου, ο αλγόριθμος αποτελείται από τρία στάδια, όπως φαίνεται στην Εικόνα 4.3.

1. Τον ορισμό της πληροφοριακής μονάδας αναφοράς, όπου ορίζεται μια σχετική με τις πληροφοριακές ανάγκες του χρήστη ιστοσελίδα, που θα αποτελέσει αφετηρία για την αναζήτηση.
2. Την αναζήτηση σχετικών πληροφοριακών μονάδων, που αποτελεί μια προσαρμοσμένη έκδοση του αλγόριθμου αποικίας μυρμηγκιών. Ο αλγόριθμος λαμβάνει υπόψη την ομοιότητα εγγράφων (αποτελεί το μέτρο αξιολόγησης της σχετικής πληροφορίας) για να αποφευχθεί η εξαντλητική διάσχιση ενός μεγάλου πλήθους υποψήφιων λύσεων. Ο αλγόριθμος είναι αναδρομικός και κάθε φορά που πραγματοποιείται σύγκληση σε μια πληροφοριακή μονάδα αυτή ορίζεται ως το νέο σημείο αφετηρίας αναζήτησης και επαναλαμβάνεται η διαδικασία.
3. Την οργάνωση των αποτελεσμάτων σε συστάδες: τα αποτελέσματα διαχωρίζονται σε συστάδες με βάση το βαθμό σχετικότητας των περιεχομένων τους.

Ο Ant-Seeker εκτελεί στοχαστική αναζήτηση για την εύρεση ιστοσελίδων στον Παγκόσμιο Ιστό με περιεχόμενο παρόμοιο με μια ιστοσελίδα αναφοράς καθώς και την ανακάλυψη των

υπερσυνδέσμων που τις ενώνουν. Ο προτεινόμενος αλγόριθμος υιοθετεί τα περισσότερα βασικά χαρακτηριστικά της οικογένειας αλγορίθμων αποικίας μυρμηγκιών. Πιο συγκεκριμένα:



Εικόνα 4.3 - Μια άποψη υψηλού επιπέδου της προτεινόμενης προσέγγισης

- νοήμονες πράκτορες, που στο πλαίσιο αυτό ονομάζονται μυρμήγκια, διατρέχουν το γράφο μετακινούμενα από κόμβο σε κόμβο, επιλέγοντας τον επόμενο σταθμό τους με βάση τα ερεθίσματα που δέχθηκαν από τον προηγούμενο.
- σε κάθε κόμβο που επισκέπτονται, οι πράκτορες αποθηκεύουν πληροφορία ανάλογη με την ποιότητα της διαδρομής. Η πληροφορία αυτή προσομοιώνει την φερομόνη που εναποθέτουν τα πραγματικά μυρμήγκια κατά τη μετακίνησή τους στο περιβάλλον.
- ποιοτικότερες διαδρομές ελκύουν περισσότερα μυρμήγκια λόγω της μεγαλύτερης συγκέντρωσης φερομόνης. Εξαιτίας αυτού, η ποσότητα της φερομόνης ενισχύεται ακόμα περισσότερο και με την πάροδο του χρόνου και κατά συνέπεια η αναζήτηση συγκλίνει προς μια λύση.

Εξαιτίας των ιδιαιτεροτήτων που παρουσιάζει το πρόβλημα αναζήτησης σχετικών ιστοσελίδων στον Παγκόσμιο Ιστό, η εισαγωγή ορισμένων τροποποιήσεων στον κλασικό αλγόριθμο, κρίνεται απαραίτητη. Πιο συγκεκριμένα: α) επειδή οι διαστάσεις του Παγκόσμιου Ιστού είναι άπειρες, η εξαντλητική αναζήτηση στον γράφο που τον αναπαριστά θα οδηγούσε σε απαγορευτικούς χρόνους εκτέλεσης, β) η αναζήτηση έχει να κάνει σχέση με την εννοιολογική σχετικότητα ιστοσελίδων και γ) ως σημείο αναφοράς της σύγκρισης είναι μια αρχική ιστοσελίδα.

Έτσι οι βασικότερες από τις επιπρόσθετες ιδιότητες που εισήχθησαν στον αλγόριθμο είναι:

- κάθε τεχνητό μυρμήγκι μπορεί να επισκεφτεί έναν μέγιστο αριθμό κόμβων
- το ποσό της φερομόνης καθορίζεται από την ομοιότητα των κόμβων.
- όλα τα τεχνητά μυρμήγκια ξεκινούν από τον κόμβο αφετηρία

Μια υψηλού επιπέδου περιγραφή του Ant-Seeker φαίνεται στον Αλγόριθμος 4.1. Ως είσοδο ο αλγόριθμος λαμβάνει μια σειρά από παραμέτρους:

- τον συνολικό αριθμό μυρμηγκιών που εκτελούν την αναζήτηση (NoA). Προφανώς όσο μεγαλύτερος είναι ο αριθμός των μυρμηγκιών τόσο ασφαλέστερα είναι τα αποτελέσματα της αναζήτησης, καθώς ένα μεγαλύτερο μέρος του γράφου θα έχει καλυφθεί. Ωστόσο,

ένας μεγάλος πληθυσμός μυρμηγκιών αυξάνει και τις απαιτήσεις σε χρόνο εκτέλεσης. Να σημειωθεί δε, ότι σύμφωνα με τα πειραματικά αποτελέσματα της επόμενης παραγράφου, η ακρίβεια δεν αυξάνει γραμμικά με τον αριθμό μυρμηγκιών που έχουν επιστρατευτεί.

- την αρχική τιμή φερομόνης (IPV) για την οποία θα πρέπει να ισορροπεί μεταξύ των ακραίων τιμών, αφού δεν πρέπει ούτε να καθοδηγεί την αναζήτηση προς τους ανεξερεύνητους κόμβους ούτε και να αποτρέπει την εξερεύνησή τους.
- το μέγιστο αριθμό κόμβων (N_{max} που μπορεί να επισκεφτεί ένα μυρμήγκι μέχρι την ολοκλήρωση της διαδρομής του. Η τιμή αυτή ουσιαστικά καθορίζει και το βάθος της αναζήτησης.

Η εκκίνηση της αναζήτησης γίνεται από τον αρχικό κόμβο-αφετηρία (d_0) ο οποίος είναι μια ιστοσελίδα σχετική με τις πληροφοριακές ανάγκες του χρήστη. Στον αλγόριθμο, υπάρχει ένας πληθυσμός από τεχνητά μυρμήγκια (NoA), καθένα από τα οποία, εκτελεί μια κίνηση από έναν κόμβο i προς έναν κόμβο j (σε κάθε επανάληψη των γραμμών 5-9). Δυνατότητα μετακίνησης από έναν κόμβο i προς έναν κόμβο j μπορεί να υπάρξει μόνο όταν δυο κόμβοι είναι άμεσα συνδεδεμένοι. Η επιλογή της επόμενης σελίδας (γραμμή 6) γίνεται μετά από μια σειρά βημάτων που περιγράφεται στην παράγραφο 4.4.1.

Με κάθε κόμβο που έχει επιλέξει ένα μυρμήγκι με βάση την τιμή της φερομόνης δημιουργείται μια διαδρομή P_t (γραμμές 3-12). Κατά τις αρχικές επαναλήψεις (γραμμές 2-27), τα μυρμήγκια αναμένεται να δημιουργούν διαδρομές με συνολικά χαμηλή ποιότητα, ωστόσο όσο προχωράνε οι επαναλήψεις οι κόμβοι που παρουσιάζουν υψηλότερη ποιότητα αυξάνουν την τιμή φερομόνης τους, οπότε και επιλέγονται με ακόμα μεγαλύτερη πιθανότητα.

Κάθε φορά που γίνεται μετακίνηση ενός μυρμηγκιού σε ένα νέο κόμβο, υπολογίζεται και η ομοιότητα του κειμένου της ιστοσελίδας του συγκεκριμένου κόμβου σε σχέση με το κείμενο του κόμβου-αναφορά, για να χρησιμοποιηθεί κατά την ανανέωση της φερομόνης. Έτσι για κάθε κόμβο ανατίθεται μια τιμή ομοιότητας η οποία δίνεται από την σχέση (4.12) όπως αναλύεται στην παράγραφο 4.4.2. Η ομοιότητα επίσης προσμετρείται στον τελικό υπολογισμό της ποιότητας του κόμβου (παράγραφος 4.4.3). Ένα ακόμα κριτήριο το οποίο παίζει ρόλο στη διαμόρφωση της ποιότητας ενός κόμβου είναι και το κατά πόσο ο συγκεκριμένος κόμβος οδηγεί σε έναν άλλο κόμβο υψηλής ποιότητας.

Η διαδικασία επιλογής κόμβων επαναλαμβάνεται μέχρι το κάθε μυρμήγκι να επισκεφτεί έναν μέγιστο αριθμό κόμβων (N_{max}). Μετά την ολοκλήρωση των μετακινήσεων όλων των μυρμηγκιών, εντοπίζεται ο καλύτερος κόμβος και ανανεώνονται οι τιμές φερομόνης των κόμβων (γραμμές 15-27). Η όλη διαδικασία αυτή επαναλαμβάνεται από έναν νέο πληθυσμό μυρμηγκιών, μέχρι να υπάρξει σύγκλιση σε κάποια συγκεκριμένη διαδρομή, δηλαδή ένα προκαθορισμένο πλήθος μυρμηγκιών

(NC_{max}) να επισκεφτεί μια διαδρομή που θα περιέχει τον κόμβο με τη μέγιστη τιμή ομοιότητας προς τον κόμβο αφετηρία.

Αλγόριθμος 4.1 - Ο αλγόριθμος Ant-Seeker

```
Είσοδος:      d0, η ιστοσελίδα αφετηρίας αναζήτησης
              NoA, πλήθος μυρμηγκιών
              IPV, αρχική τιμή φερομόνης
              NCmax, το πλήθος μυρμηγκιών προς σύγκλιση
              Nmax, το μέγιστο πλήθος ιστοσελίδων προς επίσκεψη
Έξοδος:       dbest, η περισσότερο σχετική ιστοσελίδα με τη d0

1. P = {}, t = 0, nc = 0, dbest-1 = null, dbest = null
2. ΕΠΑΝΕΛΑΒΕ
3.    ΟΣΟ (t < NoA)
4.       Pt = {}, n = 0
5.       ΟΣΟ n < Nmax
6.          dn+1 = επιλογή_επόμενης_ιστοσελίδας(dn, Pt)
7.          Υπολόγισε την ομοιότητα της dn+1 με την αρχική
8.          Πρόσθεσε τη dn+1 στο Pt
9.          n = n + 1
10.       ΤΕΛΟΣ ΕΠΑΝΑΛΗΨΗΣ
11.      Πρόσθεσε το μονοπάτι Pt στη λίστα P
12.      t = t + 1
13.   ΤΕΛΟΣ ΕΠΑΝΑΛΗΨΗΣ
14.   nc = 0
15.   ΓΙΑ ΚΑΘΕ μονοπάτι p στη λίστα P
16.      ΓΙΑ ΚΑΘΕ ιστοσελίδα d στο μονοπάτι p
17.         Ενημέρωσε τη φερομόνη της ιστοσελίδας d
18.         Υπολόγισε την ποιότητα της ιστοσελίδας d
19.      ΤΕΛΟΣ ΕΠΑΝΑΛΗΨΗΣ
20.         ΑΝ η d^p_best είναι καλύτερη από τη dbest
21.           dbest = d^p_best
22.           nc = 0
23.
24.         ΑΛΛΙΩΣ ΑΝ η dbest είναι η ίδια με την d^p_best
25.            nc = nc + 1
26.         ΤΕΛΟΣ ΑΝ
27.   ΤΕΛΟΣ ΕΠΑΝΑΛΗΨΗΣ
28.ΟΣΟ(nc < NCmax)
29.ΕΠΕΣΤΡΕΨΕ dbest
```

Ως λύση ορίζεται ο κόμβος με την μεγαλύτερη τιμή ομοιότητας με το αρχικό έγγραφο αναφοράς. Προαιρετικά, ο κόμβος με τη μεγαλύτερη ομοιότητα, επιλέγεται και στη συνέχεια

ορίζεται ως αφετηρία για νέα αναζήτηση, μέχρι ο αλγόριθμος να επιστρέψει έναν ικανοποιητικό αριθμό όμοιων εγγράφων. Στη συνέχεια μπορεί να εφαρμοστεί μια τεχνική συσταδοποίησης, στόχος της οποίας θα είναι η καλύτερη παρουσίαση των αποτελεσμάτων. Τα έγγραφα τα οποία περιέχουν πληροφοριακές μονάδες περισσότερο συσχετιζόμενες εντάσσονται στην ίδια συστάδα. Η δημιουργία τέτοιων συστάδων βασίζεται στην ανάλυση ιστογράμματος ομοιοτήτων, το οποίο είχε προταθεί αρχικά στο [59]. Για τον προσδιορισμό της ομοιότητας εγγράφων δε, μπορεί να χρησιμοποιηθεί το κλασικό χωροδιανυσματικό μοντέλο [56] [57].

Σημαντικό ρόλο στην αναζήτηση του αλγορίθμου, παίζουν οι μεταβλητές που ο αλγόριθμος λαμβάνει ως παραμέτρους από το χρήστη. Ο τρόπος με τον οποίο αυτές επηρεάζουν τα αποτελέσματα μελετήθηκε πειραματικά και παρουσιάζεται στην παράγραφο 4.5.

4.4.1 Επιλογή επόμενου κόμβου

Κάθε μυρμήγκι που βρίσκεται σε έναν κόμβο i και το οποίο δεν έχει ολοκληρώσει την περιήγησή του (δηλαδή να έχει επισκεφτεί N_{max} κόμβους), καλείται να επιλέξει τον επόμενο κόμβο j (γραμμή 6). Οι διαθέσιμες επιλογές του είναι οι κόμβοι που συνδέονται άμεσα με τον τρέχοντα κόμβο, δηλαδή οι ιστοσελίδες στις οποίες κάνουν αναφορά οι υπερσύνδεσμοι τις τρέχουσας σελίδας. Για την αποφυγή κυκλικών διαδρομών και κατ' επέκταση την παρεμπόδιση της επέκτασης της διαδρομής προς ανεξερεύνητους κόμβους, αποκλείονται οι κόμβοι οι οποίοι ήδη αποτελούν μέρος της διαδρομής (κόμβοι στη λίστα κόμβων P_t). Με αυτόν τον τρόπο επιτυγχάνεται μια συνέχεια της κίνησης στον γράφο καθώς και η δημιουργία φυσικών διαδρομών στον Παγκόσμιο Ιστό με αρχή και τέλος. Έτσι ο αποκλεισμός των προηγούμενων κόμβων (που γενικά είναι γνωστός ως προσβασιμότητα), εξασφαλίζεται ορίζοντας $n_{ij} = 1$, αν το j συνδέεται άμεσα με το i και $n_{ij} = 0$ αν το j δεν συνδέεται άμεσα με το i ή αν το j ναι μεν συνδέεται άμεσα με το i αλλά περιέχεται ήδη στη διαδρομή που έχει διανύσει ως τώρα το μυρμήγκι.

$$\eta_{ij} = \begin{cases} 1 \\ 0 \end{cases} \tag{4.12}$$

$$P_{ij} = \frac{\tau_j \eta_{ij}}{\sum_k \tau_k \eta_{ik}} \tag{4.13}$$

Όπως φαίνεται από τη σχέση (4.13), η πιθανότητα επιλογής του επόμενου κόμβου μετάβασης j (γραμμή 6) από τον τρέχον κόμβο i, εξαρτάται από τη φερομόνη τ_j που έχει αποτεθεί στον συγκεκριμένο κόμβο. Προφανώς, όσο μεγαλύτερη είναι η συγκέντρωση της φερομόνης τόσο πιο μεγάλη η πιθανότητα επιλογής του κόμβου από τα μυρμήγκια πράκτορες, ωστόσο δεν είναι αναγκαίο ότι πάντα θα προτιμάται ο κόμβος με τη μεγαλύτερη ποσότητα φερομόνης. Έτσι δίνεται στο σύστημα η δυνατότητα να εξετάζει και νέες πιθανές λύσεις. Η ανανέωση της φερομόνης γίνεται σε επόμενη φάση και πιο συγκεκριμένα στις γραμμές 15-27 ενώ περιγράφεται στην παράγραφο.

4.4.2 Ομοιότητα εγγράφων

Η ομοιότητα των εγγράφων εξαρτάται α) από τη συχνότητα εμφάνισης κοινών όρων που περιέχονται στα έγγραφα αυτά, σύμφωνα με το χωροδιανυσματικό μοντέλο [123], [124] αλλά και β) την ομοιότητα των προτάσεων στα έγγραφα, σύμφωνα με ένα τροποποιημένο μοντέλο που περιγράφεται στο [125] ώστε να λαμβάνεται υπόψη και η θέση (σημαντικότητα) των φράσεων στο κείμενο. Η ολική τιμή ομοιότητας δυο εγγράφων υπολογίζεται από τη σχέση (4.14):

$$S(d_1, d_2) = 0.5S_p(d_1, d_2) + 0.5S_t(d_1, d_2) \tag{4.14}$$

όπου d_1, d_2 είναι δύο έγγραφα και S_p είναι η ομοιότητα των προτάσεων στα έγγραφα αυτά ενώ S_t είναι η ομοιότητα των όρων. Πιο συγκεκριμένα η ομοιότητα προτάσεων υπολογίζεται από τη σχέση (4.15):

$$S_p(d_1, d_2) = \frac{\sqrt{\sum_{i=1}^{p}[g(l_i)(f_{i1}w_{i1} + f_{i2}w_{i2})]^2}}{\sum_j |s_{j1}|w_{j1} + \sum_k |s_{k2}|w_{k2}} \tag{4.15}$$

όπου p είναι ο αριθμός των κοινών φράσεων, l_i το μήκος του κοινού τμήματος των φράσεων $|s_\iota|$ το συνολικό μήκος των προτάσεων, f_i η συχνότητα εμφάνισής τους στο κείμενο των εγγράφων, w_i τα επίπεδα σημασίας των κοινών φράσεων που εντοπίστηκαν, και $g(l_i)$ είναι μια συνάρτηση που βαθμολογεί το μήκος της κοινής φράσης και δίνει υψηλότερη βαθμολογία όσο το μήκος της κοινής φράσης προσεγγίζει το μήκος της αρχικής. Η συνάρτηση $g(l_i)$ υπολογίζεται ως εξής:

$$g(l_i) = \left(\frac{|ms_i|}{|s_i|}\right)^\gamma \tag{4.16}$$

όπου s_i είναι το συνολικό μήκος της πρότασης $|ms_i|$ είναι το τμήμα της πρότασης που είναι κοινό και γ είναι ο δείκτης τεμαχισμού της πρότασης. Για γ=1 τα δυο μισά μιας πρότασης μπορούν να βρουν αντιστοιχία ανεξάρτητα το ένα από το άλλο. Ωστόσο για γ > 1 τότε ολόκληρες οι προτάσεις βαθμολογούνται υψηλότερα από τα τμήματά τους.

Παράλληλα, συνυπολογίζεται και ο βαθμός της ομοιότητας των εγγράφων με βάση τους όρους που αυτά περιέχουν. Ο υπολογισμός αυτός γίνεται σύμφωνα με το χωροδιανυσματικό μοντέλο που περιγράφτηκε στην παράγραφο 4.3.2.1. Στο μοντέλο αυτό, ο βαθμός ομοιότητας μεταξύ δυο κειμένων αντιστοιχεί στο βαθμό συσχέτισης μεταξύ των δύο διανυσμάτων που αναπαριστούν τα έγγραφα αυτά. Μέτρο του βαθμού συσχέτισης αποτελεί το συνημίτονο της γωνίας που σχηματίζεται μεταξύ των δύο διανυσμάτων. Αυτό δίνεται από τη σχέση:

$$sim(d_i, d_j) = \frac{d_i d_j}{|d_i||d_j|} = \frac{\sum_{k=1}^{t} w_{k,i}w_{k,j}}{\sqrt{\sum_{k=1}^{t} {w_{k,i}}^2}\sqrt{\sum_{k=1}^{t} {w_{k,j}}^2}} \tag{4.17}$$

όπου $w_{k,i}$ τα βάρη των όρων του κειμένου d_i και $w_{k,j}$ τα βάρη των όρων του κειμένου d_j.

Όπως αναφέρθηκε στην παράγραφο 4.3.2.1, η ανάθεση των βαρών γίνεται με βάση τον παράγοντα TF που δηλώνει την συχνότητα του όρου k και στα δυο κείμενα, καθώς και του παράγοντα IDF που δηλώνει πόση πληροφορία ο κάθε όρος αποδίδει, με άλλα λόγια αν ο όρος είναι σπάνιος ή κοινός στα κείμενα. Ωστόσο στην συγκεκριμένη περίπτωση στην ανάθεση βαρών δεν λαμβάνεται υπόψη η αντίστροφη συχνότητα εμφάνισης του κάθε όρου (παράγοντας IDF) γιατί στην ουσία έχουμε να κάνουμε με σύγκριση μεταξύ δυο εγγράφων μόνο και όχι ενός ερωτήματος και μιας συλλογής εγγράφων. Έτσι δεν μπορούν να εξαχθούν ασφαλή συμπεράσματα για το ποιοι όροι πρέπει να θεωρούνται κοινοί σε ένα τόσο μικρό δείγμα.

4.4.2.1 Υπολογισμός επιπέδων σημασίας των κοινών φράσεων

Οι ιστοσελίδες των κόμβων που επισκέπτονται τα μυρμήγκια, περιέχουν την αξιοποιήσιμη πληροφορία σε μορφή κειμένου αναμεμιγμένη με ειδικές ακολουθίες χαρακτήρων που αφορούν τον τρόπο με τον οποίο αυτή θα παρουσιάζεται στον περιηγητή (browser) του τελικού χρήστη. Αυτό είναι το λεγόμενο markup που οριοθετείται από συγκεκριμένες ετικέτες HTML.

Θεωρητικά οι ετικέτες HTML δεν φέρουν αξιοποιήσιμη πληροφορία, οπότε ένας απλός διαχωρισμός των ακολουθιών HTML από το κείμενο, ακολουθούμενος από μια καταμέτρηση των εμφανιζόμενων όρων σε αυτό θα αρκούσε για να αξιολογήσει την ομοιότητα μεταξύ δυο εγγράφων. Στην πράξη ωστόσο κάποια τμήματα του εγγράφου περιέχουν μεγαλύτερη ποσότητα αξιοποιήσιμης πληροφορίας σε σχέση με άλλα. Κατά κανόνα τα τμήματα αυτά ορίζονται από συγκεκριμένες HTML ετικέτες. Για παράδειγμα, ο τίτλος της σελίδας (οριοθετείται από την ετικέτα <Title>) είναι σημαντικότερος από μια τυχαία παράγραφο (οριοθετείται από την ετικέτα <p>) στο σώμα του εγγράφου. Για το λόγο αυτό ορισμένες HTML ετικέτες θα πρέπει να συνεκτιμούνται περισσότερο καθώς προσδίδουν διαφορετική βαρύτητα στους κειμενικούς όρους που περιλαμβάνουν.

Στην προτεινόμενη προσέγγιση τα τμήματα των εγγράφων κατατάσσονται σε τρία επίπεδα σημασίας (ΥΨΗΛΟ, ΜΕΣΑΙΟ, ΧΑΜΗΛΟ), ανάλογα με το ποσό καθαρής πληροφορίας που αναμένεται να φέρουν και τα οποία οριοθετούνται από ανάλογες HTML ετικέτες. Έτσι, για παράδειγμα, τμήματα υψηλής βαρύτητας είναι ο τίτλος, οι επικεφαλίδες καθώς και τα μεταδεδομένα. Μεσαίας σημασίας θεωρούνται τα τμήματα που εμφανίζονται έγχρωμα, ή με έντονη, υπογραμμισμένη ή πλάγια γραμματοσειρά. Τέλος, χαμηλής σημασίας θεωρείται το υπόλοιπο σώμα του κειμένου. Στον Πίνακας 4.1 περιέχονται περισσότερα τέτοια παραδείγματα.

Πίνακας 4.1 - Σπουδαιότητα διαφόρων ετικετών σε μια HTML σελίδα

Ετικέτα	**Σημασία**	**Σπουδαιότητα**
<Title>	Τίτλος	ΥΨΗΛΟ
<Meta Name=”description”>	Περιγραφή ως μεταδεδομένα	ΥΨΗΛΟ
<Meta Name=”keyword”>	Λέξεις κλειδιά ως μεταδεδομένα	ΥΨΗΛΟ

<H1>, <H2>	Επικεφαλίδα διαφορετικών επιπέδων	ΜΕΣΑΙΟ
<B>, <I>, <U>		ΜΕΣΑΙΟ
<Table>	Πίνακας	ΧΑΜΗΛΟ
<P>		ΧΑΜΗΛΟ

Έτσι λοιπόν, αφού γίνει ο διαχωρισμός της κειμενικής πληροφορίας από τις HTML ετικέτες γίνεται ο υπολογισμός των συχνοτήτων των όρων για κάθε ιστοσελίδα. Στη συνέχεια με βάση το κείμενο αυτό, τη βαρύτητά του, τους όρους και τις προτάσεις που αυτό περιέχει, γίνεται ο υπολογισμός της ομοιότητας των κόμβων όπως περιγράφτηκε προηγουμένως.

4.4.3 Υπολογισμός ποιότητας κόμβων

Για τον υπολογισμό της ποιότητας των κόμβων h(t) από ένα μυρμήγκι, λαμβάνεται υπόψη η ομοιότητά σε σύγκριση με τον αρχικό κόμβο-αναφοράς, όλων των κόμβων σε μια διαδρομή. Η λογική πίσω από αυτό είναι απλή: Αν θεωρηθεί ότι ο Παγκόσμιος Ιστός είναι μια κατανομή ομοιοτήτων εγγράφων στο χώρο, τότε αυτή η κατανομή όπως είναι αναμενόμενο, θα παρουσιάζει περιοχές με πυκνώματα και αραιώματα (υψηλές και χαμηλές τιμές ομοιότητας). Όμως είναι πιθανόν οι περιοχές υψηλής πυκνότητας να μην εντοπίζονται όλες μαζί απομονωμένες σε ένα τμήμα του χώρου. Κατά την αναζήτηση περιοχών υψηλής ποιότητας, ενδέχεται η διαδρομή να περιλαμβάνει μετάβαση σε περιοχές χαμηλής πυκνότητας. Συνεπώς, τα σημεία της περιοχής με χαμηλή τιμή ομοιότητας πρέπει να αυξάνουν την σπουδαιότητά τους όταν οδηγούν σε περιοχές υψηλής ποιότητας. Με άλλα λόγια, εάν ένας κόμβος οδηγεί σε κόμβους υψηλής ομοιότητας, θα πρέπει να επιλέγεται με μεγαλύτερη πιθανότητα ακόμα και αν ο ίδιος έχει χαμηλή τιμή ομοιότητας.

Ο υπολογισμός της ποιότητας ενός κόμβου λοιπόν, αποτελεί ουσιαστικά την ευρετική συνάρτηση που κατευθύνει την αναζήτηση και δίνεται από την σχέση:

$$h_i(t+1) = \max\left(S_j^d, S_i, h(t)\right), \text{όπου } i < j < N_{max} \tag{4.18}$$

όπου S_j^d συνάρτηση ομοιότητας του κόμβου j που αποτελεί μέρος μια διαδρομής στην οποία συμμετέχει ο κόμβος i αλλά ξεκινώντας μετά από αυτόν, και S_i η συνάρτηση ομοιότητας του κόμβου i που δίνεται από τη σχέση 4.1, με $i < j < Nmax$.

4.4.4 Ενημέρωση φερομόνης

Κατά την έναρξη του αλγόριθμου, ανατίθεται μια αρχική τιμή φερομόνης σε όλους τους κόμβους του γράφου (IPV) και αυτή μεταβάλλεται σύμφωνα με το κατά πόσο ένας κόμβος προτιμήθηκε και αποτέλεσε μέρος μιας διαδρομής ενός μυρμηγκιού. Οι κόμβοι οι οποίοι χρησιμοποιήθηκαν ως ενδιάμεσοι ή τελικοί σταθμοί στις διαδρομές των μυρμηγκιών αυξάνουν τη συγκέντρωση της φερομόνης ανάλογα με την ποιότητά τους. Η αύξηση αυτή γίνεται σύμφωνα με τη σχέση:

$$\tau_i(t+1) = \tau_i(t) + kh_i \tag{4.19}$$

όπου $t_i(t)$ η υπάρχουσα τιμή φερομόνης στον κόμβο i, h_i η ευρετική συνάρτηση που δίνεται από τη σχέση (4.18) και καθορίζει την ποιότητα του κόμβου i, ενώ k είναι ο αριθμός των μυρμηγκιών που χρησιμοποίησαν τον κόμβο i για τη δημιουργία της διαδρομής τους.

Αντίθετα, σε κάθε επανάληψη όσοι κόμβοι δεν χρησιμοποιήθηκαν καθόλου ως μέρος διαδρομής υπόκεινται σε μείωση της τιμής της φερομένης τους. Πιο συγκεκριμένα η μείωση αυτή ισοδυναμεί με μια κανονικοποίηση στο διάστημα $[0,1]$:

$$\tau_i(t+1) = \frac{\tau_i(t+1)}{\tau_{max}(t+1)} \tag{4.20}$$

Η αρχική τιμή της φερομόνης ορίζεται επίσης στο διάστημα $[0,1]$ και καθορίζει την κατεύθυνση της αναζήτησης προς ανεξερεύνητους ή όχι κόμβους. Η χρήση της κανονικοποίησης βελτιώνει την ικανότητα αναζήτησης του αλγορίθμου διότι κόμβοι χαμηλής ποιότητας αφαιρούνται από τον γράφο αναζήτησης ενώ τη θέση τους καταλαμβάνουν ανεξερεύνητοι κόμβοι.

4.5 Πειραματική αξιολόγηση

Στην παράγραφο αυτή παρουσιάζονται τα αποτελέσματα της αξιολόγησης της προτεινόμενης μεθοδολογίας. Για την αξιολόγηση εκτελέστηκαν δυο σειρές πειραμάτων. Η πρώτη σειρά αποσκοπεί στην αξιολόγηση της απόδοσης του προτεινόμενου αλγορίθμου μετά από την ολοκλήρωση τριών εκτελέσεών του. Στη δεύτερη αξιολογείται η αποτελεσματικότητα της ομαδοποίησης των επιστρεφόμενων αποτελεσμάτων. Να σημειωθεί ότι για την ανάλυση των αποτελεσμάτων απαιτήθηκε η πλήρης σάρωση των περιοχών αναζήτησης.

4.5.1 Αξιολόγηση αναζήτησης

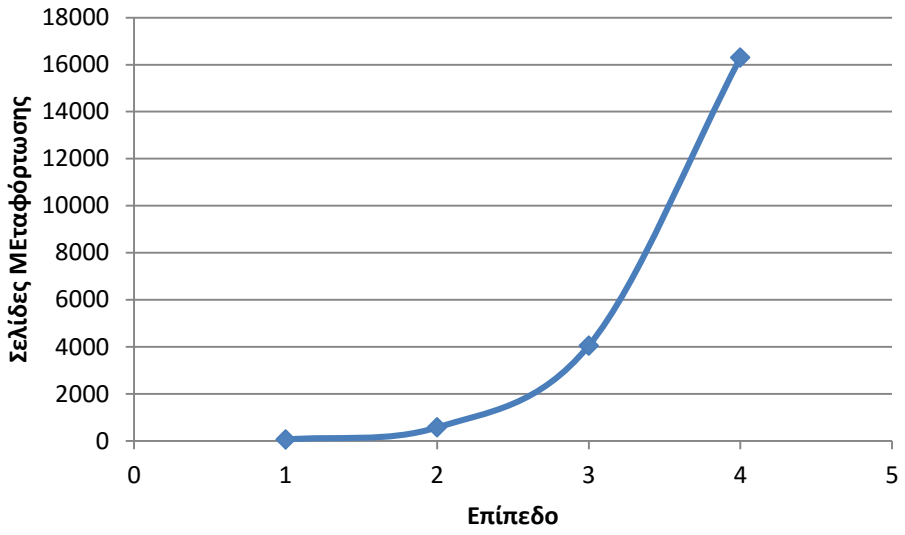
Ο προτεινόμενος αλγόριθμος αναζήτησης εκτελέστηκε με είσοδο τρεις διαφορετικές σελίδες αναφοράς που αφορούσαν ανεξάρτητα τμήματα του Παγκόσμιου Ιστού. Στην όλη διαδικασία χρησιμοποιήθηκαν μόνο ιστοσελίδες των οποίων το περιεχόμενο ήταν στην αγγλική γλώσσα. Για την ελαχιστοποίηση του χρόνου εκτέλεσης της διαδικασίας αναζήτησης και την ασφαλέστερη αξιολόγηση των αποτελεσμάτων, έγινε η μεταφόρτωση όλων των ιστοσελίδων που ανταποκρίνονται στα τμήματα του παγκόσμιου ιστού που υπάρχουν γύρω από το έγγραφο αναφοράς όπως δείχνει ο Πίνακας 4.2.

Πίνακας 4.2 - Δείγματα αξιολόγησης του αλγόριθμου

Δείγματα	**Αριθμός Ιστοσελίδων**	**Αριθμός Ιστοσελίδων προς Αναζήτηση**
1	140594	43
2	192155	34
3	175977	72

Λόγω της εκθετικής εξάπλωσης του Παγκόσμιου Ιστού, ο συνολικός αριθμός ιστοσελίδων που αναμένεται να μεταφορτωθεί για μια αναζήτηση σε βάθος 10 επιπέδων είναι απαγορευτικός. Πιο συγκεκριμένα, έχει αποδειχθεί πειραματικά ότι η εξάπλωση του Παγκόσμιου Ιστού ακολουθεί μια κατανομή α^x, όπου α είναι ο μέσος όρος υπερσυνδέσεων ανά σελίδα και x είναι το επίπεδο από μια ιστοσελίδα αναφοράς. Ενδεικτικά για έναν μέσο όρο 10 υπερσυνδέσμων ανά ιστοσελίδα ο συνολικός αριθμός ιστοσελίδων που καταφορτώνονται για τη σάρωση βάθους 10 επιπέδων είναι της τάξης των 10^{10}.

Στην Εικόνα 4.4 φαίνεται ενδεικτικά ο ρυθμός αύξησης των ιστοσελίδων που μεταφορτώθηκαν ακολουθώντας υπερσυνδέσμους με απόσταση από 1 μέχρι 4 επίπεδα από την ιστοσελίδα-αφετηρία. Στο συγκεκριμένο πείραμα ως σημείο αναφοράς θεωρήθηκε η σελίδα του ιστότοπου του ΕΜΠ.



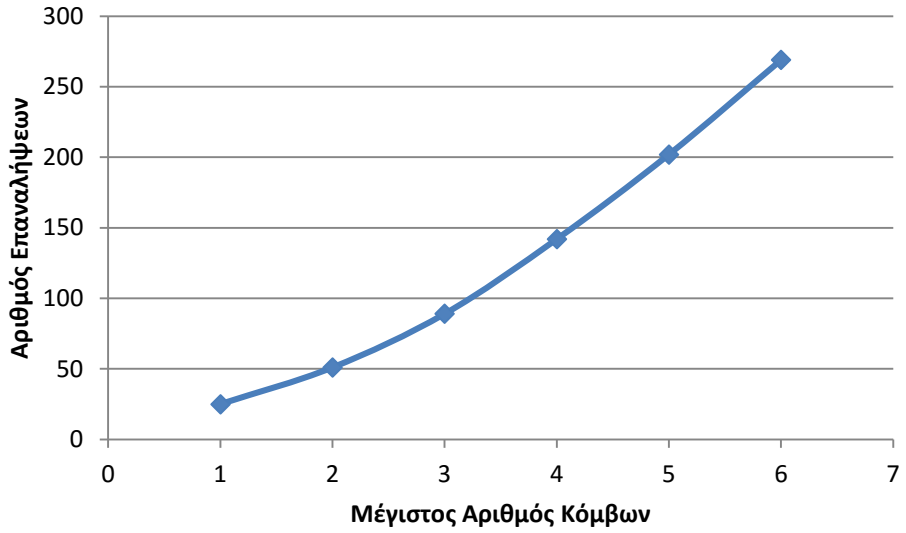
Εικόνα 4.4 - Πλήθος σελίδων ανά επίπεδο μεταφόρτωσης

Κατά την μεταφόρτωση των ιστοσελίδων, σε αρχική φάση, αποθηκεύτηκαν τα δείγματα σε μια βάση δεδομένων και έπειτα κάθε μια από αυτές υποβλήθηκαν σε μια διαδικασία προ-επεξεργασίας δυο φάσεων. Στην πρώτη φάση έγινε ο διαχωρισμός του κειμενικού περιεχομένου των ιστοσελίδων από τις HTML ετικέτες. Ταυτόχρονα σε αυτό το στάδιο, αποθηκεύτηκε και η δομή του συγκεκριμένου τμήματος του ιστού δηλαδή η περιγραφή των συνδέσεων μεταξύ κόμβων του γράφου. Με αυτόν τον τρόπο αποθηκεύτηκαν σημαντικές παράμετροι για τον αλγόριθμο όπως η δομή του γράφου, το σημείο εκκίνησης και το σημείο προορισμού. Στη δεύτερη φάση υπολογίστηκε η τιμή ομοιότητας των κειμενικών περιεχομένων των ιστοσελίδων προς το αρχικό έγγραφο αναφοράς. Αν και ο πιο αποτελεσματικός τρόπος για την πραγματοποίηση αυτής της διαδικασίας είναι η άποψη του χρήστη που θέτει το ερώτημα, η καταμέτρηση και αξιολόγηση ενός τέτοιου αριθμού εγγράφων

καθιστά αδύνατη τη συμμετοχή του ανθρώπινου παράγοντα. Έτσι για τον υπολογισμό ομοιότητας ακολουθήθηκε η αυτόματη διαδικασία που περιγράφτηκε σε προηγούμενη παράγραφο.

4.5.1.1 Το μέγιστο βάθος αναζήτησης N_{max}

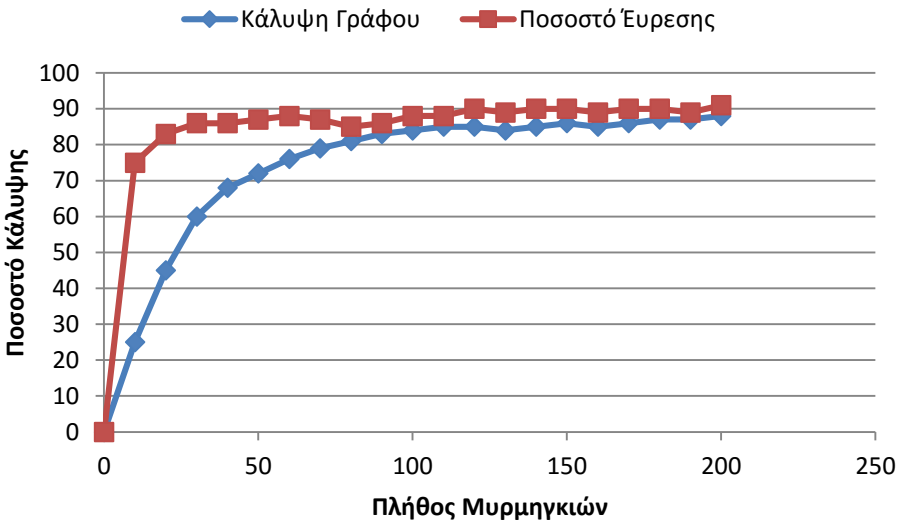
Θεωρητικά, όσο μεγαλύτερος είναι ο αριθμός των κόμβων που ένα μυρμήγκι πρέπει να επισκεφτεί κατά το σχηματισμό μιας διαδρομής, τόσο περισσότερος χρόνος απαιτείται για τη σύγκλιση του αλγορίθμου. Σύμφωνα με τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν αυτός ο ισχυρισμός φαίνεται να επαληθεύεται. Στην Εικόνα 4.5 μπορεί να παρατηρηθεί πως η αύξηση του αριθμού των επαναλήψεων είναι περίπου ανάλογη με το μήκος της διαδρομής. Να σημειωθεί ότι στα πειράματα που εκτελέστηκαν χρησιμοποιήθηκε η τιμή 3 για την συγκεκριμένη μεταβλητή.



Εικόνα 4.5 - Μέσος αριθμός επαναλήψεων ανά μήκος διαδρομής

4.5.1.2 Το πλήθος των μυρμηγκιών NoA

Θεωρητικά το πλήθος των μυρμηγκιών που εκτελούν την αναζήτηση επηρεάζουν την ποιότητα της λύσης (όσο περισσότερα τόσο καλύτερη η λύση) καθώς και το χρόνο που απαιτείται για να ολοκληρωθεί η εκτέλεση του αλγορίθμου (όσο περισσότερα τόσο μεγαλύτερος ο χρόνος εκτέλεσης). Ωστόσο, όπως φαίνεται στην Εικόνα 4.6, το πλήθος των μυρμηγκιών δεν παίζει σημαντικό ρόλο στην ποιότητα αναζήτησης. Ακόμα και για μικρές τιμές της συγκεκριμένης παραμέτρου (5-20) ο αλγόριθμος επιτυγχάνει ικανοποιητική απόδοση (70%-80%) ενώ το ποσοστό κάλυψης τους γράφου παραμένει σχετικά χαμηλό (20%-60%).



Εικόνα 4.6 - Διάγραμμα ποσοστού κάλυψης και εύρεσης λύσεων ανά αριθμό μυρμηγκιών

4.5.1.3 Αρχική τιμή φερομόνης IPV

Η αρχική τιμή φερομόνης που εναποτίθεται στους κόμβους που εισέρχονται στο χώρο αναζήτησης αποδεικνύεται ότι επηρεάζει την ταχύτητα σύγκλισης καθώς και την ποιότητα των αποτελεσμάτων. Πιο συγκεκριμένα, όσο πιο μικρή είναι η τιμή της μεταβλητής αυτής, τόσο πιο γρήγορα ο αλγόριθμος συγκλίνει σε μια λύση αλλά παράγονται κακής ποιότητας λύσεις. Από την άλλη πλευρά μεγάλες τιμές δεν βελτιώνουν θεαματικά την ποιότητα της αναζήτησης ενώ ταυτόχρονα μειώνουν την ταχύτητα σύγκλισης (πολύ μεγάλος χρόνος απόκρισης του αλγορίθμου). Αυτό διότι η λύση εξαρτάται και από τις αρχικές διαδρομές των μυρμηγκιών οι οποίες είναι τυχαίες. Δεδομένου ότι η ποιότητα του κόμβου ορίζεται από την εξίσωση ομοιότητας που δίνεται από τη σχέση (4.18) και κυμαίνεται στο διάστημα [0,1], η ιδανική τιμή για αυτή την παράμετρο είναι τέτοια που προσδίδει ουδέτερη συμπεριφορά σε κάθε νέο κόμβο. Για τα επόμενα πειράματα επιλέχθηκε η τιμή 0.4.

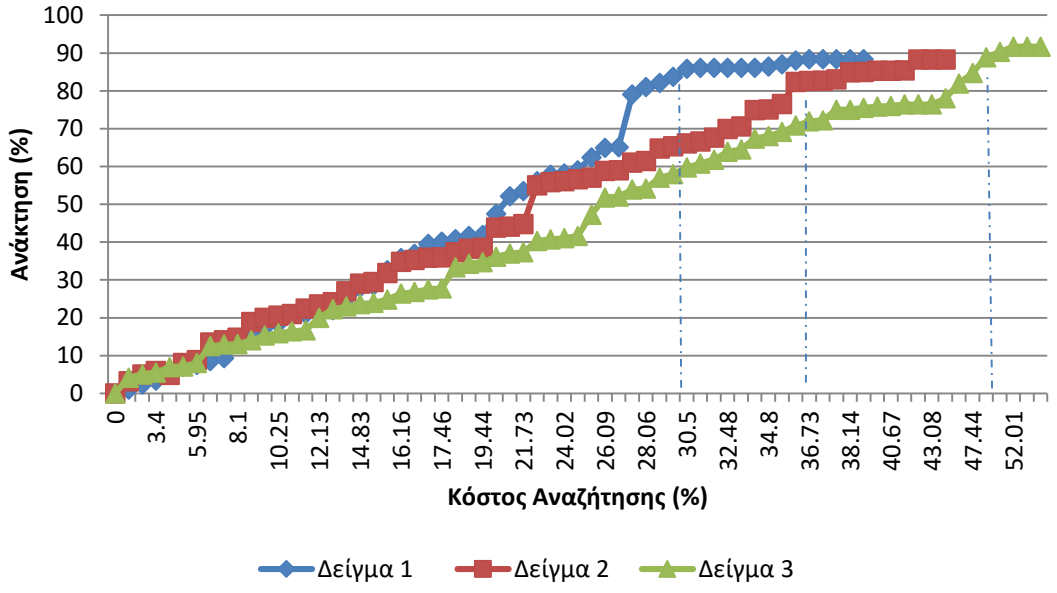
4.5.2 Αποτελέσματα Πειραματικών Μετρήσεων

Στον Πίνακας 4.4, τον Πίνακας 4.5 και τον Πίνακας 4.6, παρουσιάζονται τα πειραματικά αποτελέσματα αναζήτησης του αλγορίθμου. Για όλα τα πειράματα η τιμές των παραμέτρων που επιλέχθηκαν είναι NoA=10, Μέγιστος αριθμός κόμβων=3, NC=100 και Αρχική τιμή φερομόνης=0.4. Για κάθε σύνολο σελίδων υπολογίστηκε ο αριθμός των λύσεων που επέστρεψε ο αλγόριθμος, το κόστος εξερεύνησης, ο αριθμός των σχετικών προς το ερώτημα αποτελεσμάτων και ένας δείκτης αξιολόγησης της ποιότητας των λύσεων.

Πίνακας 4.3 - Παράμετροι αλγόριθμου για τα πειράματα

Δείγμα	Αρ. Σελίδων	Αρ. Μυρμηγκιών	N_{max}	NC	Σχετικές σελίδες
1	140594	10	3	100	43
2	192155	10	3	100	34
3	175997	10	3	100	72

Όπως αναφέρθηκε στην προηγούμενη ενότητα σε κάθε εφαρμογή του αλγορίθμου προστίθεται και ένα αποτέλεσμα είτε είναι σχετικό είτε άσχετο με το αρχικό σύνολο των προς αναζήτηση σελίδων. Με άλλα λόγια, για κάθε αναζήτηση το σύνολο επιστρεφόμενων αποτελεσμάτων είναι ίσο με τον αριθμό των επαναληπτικών εφαρμογών του αλγορίθμου. Αυτό ουσιαστικά μειώνει την ποιότητα των αποτελεσμάτων (μόνο το 40% των αποτελεσμάτων είναι σχετικό προς το αρχικό ερώτημα) αλλά επιτρέπει την περαιτέρω εξερεύνηση του γράφου όταν δε βρεθεί σχετική πληροφορία. Αυτό είναι σημαντικό ιδιαίτερα σε περιπτώσεις που ο αλγόριθμος δεν καταφέρει να συγκλίνει σε σχετική προς το ερώτημα λύση στα αρχικά στάδια της αναζήτησης. Με άλλα λόγια με κόστος την ελάττωση της ποιότητας δίνεται η δυνατότητα στον αλγόριθμο να αναζητήσει περαιτέρω λύσεις.



Εικόνα 4.7- Ποσοστό ανάκτησης και ποιότητας με την αύξηση του κόστους για τα τρία δείγματα

Παρατηρώντας την Εικόνα 4.7, προκύπτει σαν συμπέρασμα ότι με κόστος αναζήτησης περίπου 30% με 50 %, επιτυγχάνεται η ανάκτηση του 85 με 90% των εγγράφων. Αυτό σημαίνει ότι ο αλγόριθμος χρειάζεται να επισκεφτεί το 35% του γράφου μέχρι το προκαθορισμένο βάθος αναζήτησης για να ανακτήσει το 80% περίπου των σχετικών με το αρχικό εγγράφων.

Πίνακας 4.4- Αποτελέσματα εφαρμογής αλγορίθμου στο πρώτο δείγμα εγγράφων

Αρ. Εφαρμογών	Κόστος Αναζήτησης	Ποσοστό Κόστους (%)	Αριθμός Σχετικών	Ποσοστό Ανάκτησης	Ποσοστό Ποιότητας

Αλγορίθμου			Σελίδων	(%)	(%)
0	0	0,00	0	0,00	0,00
5	4563	3,25	1	2,33	20,00
10	7659	5,45	3	6,98	30,00
15	11243	8,00	4	9,30	26,67
20	12732	9,06	7	16,28	35,00
25	16023	11,40	9	20,93	36,00
30	19483	13,86	10	23,26	33,33
35	21761	15,48	14	32,56	40,00
40	24367	17,33	17	39,53	42,50
45	27337	19,44	18	41,86	40,00
50	30554	21,73	23	53,49	46,00
55	33772	24,02	25	58,14	45,45
60	36990	26,31	28	65,12	46,67
65	37813	26,90	32	74,42	49,23
70	39208	27,89	34	79,07	48,57
75	41716	29,67	36	83,72	48,00
80	43426	30,89	37	86,05	46,25
85	46238	32,89	37	86,05	43,53
90	48644	34,60	37	86,05	41,11
95	51634	36,73	38	88,37	40,00
100	54367	38,67	38	88,37	38,00

Πίνακας 4.5 - Αποτελέσματα εφαρμογής του αλγορίθμου στο δεύτερο δείγμα εγγράφων

Αρ. Εφαρμογών Αλγορίθμου	Κόστος Αναζήτησης	Ποσοστό Κόστους (%)	Αριθμός Σχετικών Σελίδων	Ποσοστό Ανάκτησης (%)	Ποσοστό Ποιότητας (%)
0	0	0,00	0	0,00	0,00
5	6535	3,40	2	5,88	40,00
10	11436	5,95	3	8,82	30,00
15	15563	8,10	5	14,71	33,33
20	19689	10,25	7	20,59	35,00
25	23311	12,13	8	23,53	32,00
30	28687	14,93	10	29,41	33,33
35	32144	16,73	12	35,29	34,29
40	37268	19,39	13	38,24	32,50
45	41623	21,66	15	44,12	33,33
50	45835	23,85	19	55,88	38,00
55	50130	26,09	20	58,82	36,36
60	56210	29,25	22	64,71	36,67
65	59644	31,04	23	67,65	35,38
70	63192	32,89	24	70,59	34,29
75	67895	35,33	26	76,47	34,67
80	69341	36,09	28	82,35	35,00
85	74868	38,96	29	85,29	34,12
90	78156	40,67	29	85,29	32,22
95	79929	41,60	30	88,24	31,58
100	84367	43,91	30	88,24	30,00

Πίνακας 4.6 - Αποτελέσματα εφαρμογής του αλγορίθμου στο τρίτο σύνολο εγγράφων

Αρ. Εφαρμογών Αλγορίθμου	Κόστος Αναζήτησης	Ποσοστό Κόστους (%)	Αριθμός Σχετικών Σελίδων	Ποσοστό Ανάκτησης (%)	Ποσοστό Ποιότητας (%)
0	0	0,00	0	0,00	0,00
5	5134	2,92	3	4,17	60,00
10	9346	5,31	5	6,94	50,00
15	13689	7,78	9	12,50	60,00
20	17235	9,79	11	15,28	55,00
25	20463	11,63	12	16,67	48,00
30	22674	12,88	16	22,22	53,33
35	26107	14,83	17	23,61	48,57
40	28438	16,16	19	26,39	47,50
45	30722	17,46	20	27,78	44,44
50	32283	18,34	24	33,33	48,00
55	36641	20,82	26	36,11	47,27
60	40762	23,16	29	40,28	48,33
65	42577	24,19	30	41,67	46,15
70	44961	25,55	34	47,22	48,57
75	49388	28,06	39	54,17	52,00
80	53676	30,50	43	59,72	53,75
85	57169	32,48	46	63,89	54,12
90	61250	34,80	49	68,06	54,44
95	64934	36,89	52	72,22	54,74
100	66599	37,84	54	75,00	54,00
105	67124	38,14	54	75,00	51,43
110	71604	40,68	55	76,39	50,00
115	75823	43,08	55	76,39	47,83
120	79368	45,10	59	81,94	49,17
125	83492	47,44	61	84,72	48,80
130	86082	48,91	64	88,89	49,23
135	89562	50,89	65	90,28	48,15
140	91534	52,01	66	91,67	47,14
145	93627	53,20	66	91,67	45,52
150	94311	53,59	66	91,67	44,00

4.5.3 Εφαρμογή με τη χρήση τεχνικών ομαδοποίησης

Όπως προαναφέρθηκε η χρήση της συσταδοποίησης αποσκοπεί στην βελτίωση της ποιότητας αποτελεσμάτων της αναζήτησης. Διαθέτοντας ένα σύνολο εγγράφων η ομαδοποίηση σε συστάδες εγγράφων με υψηλό βαθμό συνοχής αποκόπτει τα αποτελέσματα των οποίων η σχετικότητα με το αρχικό ερώτημα είναι χαμηλή.

Στον Πίνακας 4.7 παρουσιάζονται τα αποτελέσματα της συσταδοποίησης στα 3 σύνολα αξιολόγησης του αλγορίθμου. Η εφαρμογή της τεχνικής συσταδοποίησης στα επιστρεφόμενα αποτελέσματα έχει ως συνέπεια τη σημαντική βελτίωση της ποιότητας των επιστρεφόμενων σελίδων (από 30-50% στο 80-90%) όσον αφορά τη σχετικότητα τους με την αρχική. Το κόστος είναι η μικρή μείωση του ποσού ανάκτησης (κατά 2% με 5%) των σελίδων.

Πίνακας 4.7 - Απόδοση του συστήματος με χρήση τεχνικών συσταδοποίησης

Πείραμα	Σχετικές Σελίδες	Συνολικός Αριθμός	Βρέθηκαν	Ορισμένες από το σύστημα ως σχετικές	Σωστά τοποθετημένες	Ανάκτηση	Ακρίβεια
1	43	80	37	42	35	81.4	83.3
2	34	85	29	32	26	76.4	81.2
3	72	135	65	70	65	90.2	92.8

Στο σημείο αυτό θα πρέπει να σημειωθεί ότι για τον υπολογισμό της ομοιότητας των εγγράφων, κατά τη διαδικασία συσταδοποίησης χρησιμοποιείται το χωροδιανυσματικό μοντέλο [124] και όχι η κλασική προσέγγιση που χρησιμοποιείται από τον αλγόριθμο αναζήτησης. Έτσι, εν μέρη, αυτή η βελτίωση στην ακρίβεια οφείλεται στον διαφορετικό τρόπο υπολογισμού της ομοιότητας εγγράφων. Θα πρέπει δε να ξεκαθαριστεί ότι κατά τη διαδικασία της αναζήτησης δεν είναι δυνατό να επιστρατευτεί η ίδια συνάρτηση καθώς δεν είναι γνωστό εκ των προτέρων το σύνολο της συλλογής των εγγράφων που απαιτείται από αυτή τη συνάρτηση. Αναζητήθηκε λοιπόν, μια συνάρτηση που υπολογίζει την ομοιότητα ανά ζεύγος εγγράφων. Στον Πίνακας 4.8, αποτυπώνονται τα αποτελέσματα εφαρμογής της μεθοδολογίας για 6 τυχαία ερωτήματα στον παγκόσμιο ιστό.

Πίνακας 4.8 - Απόδοση του συστήματος σε τυχαία ερωτήματα στον Παγκόσμιο Ιστό

Πείραμα	**Σχετικές Σελίδες**	**Ορισμένες από το σύστημα**	**Σωστά τοποθετημένες**	**Ακρίβεια**	**Ανάκτηση**
1	32	37	28	87.50	75.6
2	40	56	32	80.00	57.14
3	21	25	19	90.48	76.00
4	10	8	80	80.00	100.00
5	17	17	15	90.48	88.24
6	36	42	33	80.00	78.57

4.6 Συμπεράσματα

Το μέγεθος των δειγμάτων ήταν της τάξης των 200.000. Η ορθή αξιολόγηση του αλγορίθμου προαπαιτεί τον πλήρη καθορισμό των μελών δείγματος, όσον αφορά την σχετικότητά τους με τον κόμβο αναφοράς. Η κατάταξη του συνόλου του δείγματος με βάση την ομοιότητα, ναι μεν μπορεί να παρέχει μια εκτίμηση της σχετικότητας των εγγράφων και κατά συνέπεια ένα μέτρο ταξινόμησης, αλλά παραμένει να είναι μια μηχανική μέθοδος κατάταξης και δεν μπορεί να αντικαταστήσει τον παράγοντα άνθρωπο. Το μέγεθος των δειγμάτων είναι αρκετά μεγάλο για να κατηγοριοποιηθεί από ανθρώπινο χέρι. Σε τέτοιου είδους περιπτώσεις θα μπορούσαν να χρησιμοποιηθούν τεχνικές μηχανικής μάθησης όπως είναι τα νευρωνικά δίκτυα [126], [127], στις περιπτώσεις αυτές όμως απαιτείται ένα σύνολο έτοιμων και κατηγοριοποιημένων εγγράφων για την εξόρυξη των σχετικών εγγράφων.

Ένας δεύτερος περιορισμός που αποτελεί ουσιαστικά αποτέλεσμα του προηγούμενου περιορισμού, είναι η επικάλυψη μεταξύ των αναζητήσεων. Ο αλγόριθμος περιλαμβάνει έναν μηχανισμό αποτροπής επικαλυπτόμενων αναζητήσεων για την αποφυγή δημιουργίας κυκλικών διαδρομών αναζήτησης. Ωστόσο σε ένα περιορισμένο τμήμα του ιστού η απαγόρευση κίνησης προς τα πίσω θα προκαλούσε τερματισμό της αναζήτησης σε ελάχιστα μόλις βήματα (συνολικά 2 με 5 αναζητήσεις, ανά δείγμα). Για αυτό το λόγο ο μόνος περιορισμός που ορίστηκε για τη δημιουργία διαδρομών, ήταν η εμπόδιση προσθήκης κόμβου, που ανήκει στο τρέχον σύνολο των λύσεων του αλγορίθμου.

Ιδιαίτερο ενδιαφέρον παρουσιάζει και η συμπεριφορά του αλγορίθμου στα οριακά σημεία του γράφου αναζήτησης. Ως οριακά σημεία ορίζονται οι κόμβοι που ανήκουν στο τελευταίο επίπεδο μεταφόρτωσης. Η συμμετοχή αυτών των κόμβων στο γράφο κρίνεται ελλιπής διότι το σύνολο των υπερσυνδέσμων οδηγούν σε ιστοτόπους έξω από το χώρο αναζήτησης. Οπότε κάθε φορά που ο αλγόριθμος δημιουργούσε διαδρομές, πολλές από αυτές τερματιζόντουσαν σε κόμβους των ορίων. Αυτό οδηγούσε τον αλγόριθμο καταστάσεις αδράνειας (stagnant).

Ένα σημαντικό πλεονέκτημα σε σχέση με τις κλασσικές μεθόδους αναζήτησης και κατηγοριοποίησης είναι το γεγονός ότι ο προτεινόμενος αλγόριθμος δεν απαιτεί την εξαντλητική κάλυψη μιας περιοχής. Για παράδειγμα, στα πειράματα που εκτελέστηκαν το ποσοστό κάλυψης ανέρχεται μόλις στο 40%. Όπως είναι αναμενόμενο αυτό έχει κόστος στο ποσοστό των ανακτηθέντων εγγράφων. Πιο συγκεκριμένα, το ποσοστό αυτό περιορίζεται σε στο 80%, το οποίο όμως κρίνεται ικανοποιητικό

Σε αντίθεση με τις περισσότερες τεχνικές αναζήτησης ο αλγόριθμος παρέχει τη δυνατότητα χρήσης κειμενικού ερωτήματος εκτός από ερωτήματα που δομούνται από αποκλειστικά ένα έγγραφο αναφοράς. Τα ερωτήματα από όρους δε, δυνητικά οδηγούν σε ποιοτικότερες αναζητήσεις. Προφανώς, η αναζήτηση με όρους έχει το πλεονέκτημα ότι είναι σύντομη. Φυσικά το ερώτημα αναζήτησης που δομείται από τους όρους θα πρέπει να είναι περιεκτικό και ακριβές ώστε να επιστραφούν πραγματικά σχετικά αποτελέσματα. Η ταυτόχρονη χρήση ενός εγγράφου αναφοράς για την αναζήτηση βελτιώνει την ποιότητα των επιστρεφόμενων αποτελεσμάτων.

Το ποσοστό ανάκτησης που επιτυγχάνεται από τον αλγόριθμο αναζήτησης σε σχέση με το ποσοστό κάλυψης είναι ικανοποιητικό. Ωστόσο καθιστά την εφαρμογή της μεθοδολογίας απαγορευτική για αναζητήσεις σε μεγάλη κλίμακα. Για παράδειγμα, στο δεύτερο σύνολο πειραμάτων, για την επιστροφή 100 αποτελεσμάτων απαιτήθηκε χρόνος μεγαλύτερος της μίας ώρας.

Κεφάλαιο 5
Επίλογος

5.1 Επαγωγή κανόνων κατηγοριοποίησης με το μοντέλο Απεικόνισης/Μείωσης

Στο Κεφάλαιο 2 της διατριβής αυτής, παρουσιάστηκε μια προσέγγιση επαγωγής κανόνων κατηγοριοποίησης βασισμένη στο μοντέλο Απεικόνισης/Μείωσης. Το πρώτο βήμα σε αυτή την προσέγγιση είναι η διακριτοποίηση των αριθμητικών ιδιοτήτων του συνόλου που γίνεται επίσης με έναν αλγόριθμο βασισμένο στο ίδιο μοντέλο. Η προτεινόμενη προσέγγιση της παραγράφου 2.4.1, αποτελεί μια απλή και κλιμακούμενη λύση η οποία θυσιάζει λίγη από την ακρίβεια έναντι του υπολογιστικού και επικοινωνιακού κόστους. Η διακριτοποίηση μεγάλων συνόλων δεδομένων με την τεχνική Απεικόνισης/Μείωσης γενικότερα δεν είναι εύκολη διαδικασία και κατά συνέπεια χρειάζεται περεταίρω έρευνα προς αυτή την κατεύθυνση από την οποία θα μπορούσαν να προκύψουν αποδοτικές και κλιμακούμενες μέθοδοι διακριτοποίησης.

Ο αλγόριθμος επαγωγής κανόνων βασίζεται στην κλασική προσέγγιση κάλυψης και πραγματοποιεί εξαντλητική αναζήτηση για την εύρεση του βέλτιστου κανόνα σε μια επανάληψη. Η εξαντλητική αναζήτηση βέλτιστης λύσης εισάγει υψηλό υπολογιστικό φορτίο, επηρεάζοντας τη χρονική απόκριση μιας εργασίας Απεικόνισης, ωστόσο η παράμετρος αυτή ρυθμίζεται από το χρήστη. Και οι δυο προσεγγίσεις (διακριτοποίησης και επαγωγής κανόνων) σχεδιάστηκαν λαμβάνοντας υπόψη τα ιδιαίτερα χαρακτηριστικά του μοντέλου προγραμματισμού Απεικόνισης/Μείωσης. Απεδείχθη ότι με τη χρήση δυο απλών έργων Απεικόνισης/Μείωσης είναι δυνατή η αναζήτηση στο χώρο των ιδιοτήτων-τιμών και η κατασκευή κανόνων αναζητώντας στους συνδυασμούς τους, με ένα τρόπο κλιμακωτό ως προς το μέγεθος του συνόλου παραδειγμάτων. Η εύρεση του καλύτερου κανόνα ωστόσο, έχει σαν αποτέλεσμα ο αλγόριθμος να είναι ακατάλληλος σε σύνολα δεδομένων με πολλές ιδιότητες. Για το λόγο αυτό ένα σημαντικό θέμα που πρέπει να διερευνηθεί είναι η απόδοση του αλγόριθμου στην αναζήτηση του χώρου κανόνων με σύνολα δεδομένων μεγάλων (10^3) ή πολύ μεγάλων (10^6) διαστάσεων.

Η προσέγγιση επαγωγής κανόνων, αξιολογήθηκε από τρεις προοπτικές: την ακρίβεια του μοντέλου κατηγοριοποίησης, το επικοινωνιακό κόστος και την παράλληλη επίδοση. Τα αποτελέσματα δείχνουν ότι η προσέγγιση πετυχαίνει ακρίβεια συγκρίσιμη με αυτή που πετυχαίνουν γνωστοί αλγόριθμοι επαγωγής κανόνων και ότι κλιμακώνεται με την αύξηση των παραδειγμάτων. Κατά συνέπεια μπορεί να φανεί χρήσιμη για ένα μεγάλο εύρος εφαρμογών.

5.2 Ανάλυση διεπαφών σε μεγάλη κλίμακα και κατηγοριοποίησή τους ως προς τη λειτουργία τους

Στο Κεφάλαιο 3, μετά από μια εισαγωγή στο πρόβλημα της αναγνώρισης των διεπαφών αναζήτησης στον Παγκόσμιο Ιστό, έγινε μια ανάλυση του συνόλου δεδομένων Yahoo L11 α) με την τεχνική Απεικόνισης/Μείωσης για τη μετατροπή των διεπαφών από HTML μορφή σε διανυσματική μορφή και β) με το Apache Hive για την εξαγωγή συμπερασμάτων για τα χαρακτηριστικά τους. Μια τέτοια προσέγγιση είναι εφαρμόσιμη στην περίπτωση των μηχανών αναζήτησης γενικού σκοπού, όπου συνηθίζεται οι ιστοσελίδες να προσκομίζονται αδιακρίτως σε μια κεντρική τοποθεσία και να αναλύονται εκεί σε δεύτερο βήμα. Τα συμπεράσματα που παρουσιάζονται σχετικά με τις διεπαφές που εξήχθησαν από το σύνολο αυτό που αφορούν τις κατανομές συχνοτήτων εμφάνισης των διαφόρων πεδίων εισόδου τους, τις μεταξύ τους συσχετίσεις, αλλά και τα κειμενικά χαρακτηριστικά που μπορούν να χρησιμοποιηθούν σε μια κατηγοριοποίηση ανεξαρτήτως της γλώσσας στην οποία είναι γραμμένη η διεπαφή.

Πιο συγκεκριμένα στο σύνολο δεδομένων Yahoo L11, βρέθηκε ότι το δείγμα δεν στοχεύει κάποιες συγκεκριμένες θεματικές κατηγορίες, ενώ ταυτόχρονα υπάρχει ποικιλία στις εμφανιζόμενες φυσικές γλώσσες. Στο δείγμα φαίνεται ότι η πλειονότητα των ιστοσελίδων περιέχουν μια ή δυο διεπαφές, ενώ από τα στοιχεία εισόδου που αυτές περιέχουν, τα στοιχεία select, option, hidden, text και submit εμφανίζονται τουλάχιστον στις μισές διεπαφές, ενώ τα περισσότερο σπάνια είναι τα στοιχεία file, optgroup, legend και reset. Σύμφωνα με το σύνολο δεδομένων Yahoo L11, δεν υπάρχει γραμμική συσχέτιση μεταξύ των πεδίων κάτι που σημαίνει ότι δεν είναι δυνατή η πρόβλεψη της τιμής ενός στοιχείου εισόδου από την τιμή του άλλου. Τέλος από την καταμέτρηση των λέξεων στις διεπαφές φάνηκε ότι οι λέξεις get και post που αποτελούν τιμές της ιδιότητας method μιας διεπαφής, καθώς και οι λέξεις search και mail εμφανίζονται στις διεπαφές ανεξάρτητα από τη φυσική γλώσσα στην οποία είναι γραμμένη. Αυτό σημαίνει ότι μπορούν να χρησιμοποιηθούν σαν ιδιότητες για τη διαδικασία της κατηγοριοποίησης ως προς τη λειτουργία των διεπαφών.

Στη συνέχεια χρησιμοποιείται η προσέγγιση για την επαγωγή κανόνων κατηγοριοποίησης που παρουσιάστηκε στο Κεφάλαιο 2 πάνω σε ένα δείγμα από τις διεπαφές που εξήχθησαν από το σύνολο Yahoo L11 και ένα δεύτερο δείγμα από ένα σύνολο ιστοσελίδων με διεπαφές γνωστό ως TEL-8. Οι ιδιότητες που χρησιμοποιήθηκαν προέκυψαν από την ανάλυση του Yahoo L11 με στόχο να είναι ανεξάρτητες της φυσικής γλώσσας στην οποία είναι γραμμένες οι διεπαφές. Μετά από έναν εμπειρικό προσδιορισμό των παραμέτρων εισόδου, έγινε μια πειραματική μελέτη για την επίδοση τόσο της προτεινόμενης προσέγγισης όσο και άλλων προσεγγίσεων που έχουν προταθεί στη βιβλιογραφία και τα αποτελέσματα δείχνουν ότι η ακρίβεια κατηγοριοποίησης των κανόνων που παρήχθησαν από την προτεινόμενη προσέγγιση υπερέχουν των υπόλοιπων. Αυτό οφείλεται αφενός στην εξαντλητική αναζήτηση που πραγματοποιεί η προτεινόμενη προσέγγιση για την εύρεση του καλύτερου κανόνα σε

κάθε επανάληψη, αλλά αφετέρου και στην επιλογή των βέλτιστων παραμέτρων εισόδου για το συγκεκριμένο πρόβλημα. Οι κανόνες που παρήχθησαν είναι σύντομοι σε μήκος και αρκετά γενικοί για να μπορούν να εφαρμοστούν επιτυχώς για την κατηγοριοποίηση διεπαφών ως προς τη λειτουργία τους.

5.3 Αναζήτηση πληροφορίας στον Παγκόσμιο Ιστό με έναν αλγόριθμο αποικίας μυρμηγκιών

Στο Κεφάλαιο 4 περιγράφτηκε μια μεθοδολογία αναζήτησης πληροφορίας στον Παγκόσμιο Ιστό που βασίζεται στον κλασικό αλγόριθμο αποικίας μυρμηγκιών. Η προτεινόμενη προσέγγιση έχει τη δυνατότητα να εντοπίζει συναφείς πληροφοριακές μονάδες δρομολογώντας την αναζήτηση πληροφορίας στην άμεση γειτονική περιοχή μιας αρχικής πληροφοριακής μονάδας που δίνεται σαν είσοδος. Η δρομολόγηση της αναζήτησης, πραγματοποιείται στοχαστικά συνδυάζοντας τεχνικές ανάκτησης που βασίζονται στην ομοιότητα εγγράφων και τεχνικών προσομοίωσης του τρόπου επικοινωνίας των μυρμηγκιών.

Συνοπτικά, ο αλγόριθμος δεδομένης μιας αρχικής ιστοσελίδας αφήνει ένα πλήθος μυρμηγκιών να επισκεφτεί τις γειτονικές ιστοσελίδες, όπως αυτές ορίζονται από τους υπερσυνδέσμους που περιέχουν και να εκτιμήσει την ποιότητα της καθεμιάς ανάλογα με το βαθμό ομοιότητάς τους με την αρχική ιστοσελίδα και την ποσότητα φερομόνης που της αντιστοιχεί. Η επίσκεψη γίνεται βάσει μιας πιθανότητας που υπολογίζεται συναρτήσει αυτών των παραγόντων. Η διαδικασία επαναλαμβάνεται μέχρι ένα συγκεκριμένο πλήθος μυρμηγκιών να συγκλίνει στον καλύτερο κόμβο, με κάθε επανάληψη να ανανεώνει τη φερομόνη ανάλογα με την ποιότητα κάθε κόμβου.

Ένα σημαντικό πλεονέκτημα σε σχέση με τις κλασσικές μεθόδους αναζήτησης και κατηγοριοποίησης είναι το γεγονός ότι ο προτεινόμενος αλγόριθμος δεν απαιτεί την εξαντλητική κάλυψη μιας περιοχής αναζήτησης. Για παράδειγμα, στα πειράματα που εκτελέστηκαν το ποσοστό κάλυψης ανέρχεται μόλις στο 40% του συνολικού χώρου αναζήτησης. Όπως είναι αναμενόμενο αυτό έχει επίπτωση στο ποσοστό των ανακτηθέντων εγγράφων. Πιο συγκεκριμένα, το ποσοστό αυτό περιορίζεται σε στο 80%, το οποίο όμως κρίνεται ικανοποιητικό. Το ποσοστό ανάκτησης που επιτυγχάνεται από τον αλγόριθμο αναζήτησης σε σχέση με το ποσοστό κάλυψης κρίνεται επίσης ικανοποιητικό. Ωστόσο καθιστά την εφαρμογή της μεθοδολογίας απαγορευτική για αναζητήσεις σε μεγάλη κλίμακα. Για παράδειγμα, στο δεύτερο σύνολο πειραμάτων, για την επιστροφή 100 αποτελεσμάτων απαιτήθηκε χρόνος μεγαλύτερος της μίας ώρας.

5.4 Μελλοντική εργασία

Κατά την εκπόνηση της διατριβής αυτής, προέκυψαν ορισμένες ενδιαφέρουσες κατευθύνσεις για μελλοντική έρευνα.

Ο αλγόριθμος επαγωγής κανόνων που παρουσιάστηκε στο Κεφάλαιο 2 βασίζεται στην κλασική προσέγγιση κάλυψης, κάτι που σημαίνει ότι είναι επαναληπτικός. Το πλήθος των επαναλήψεων εξαρτάται από το σύνολο εκπαίδευσης. Έτσι υπάρχουν περιπτώσεις όπου θα χρειαστούν αρκετές επαναλήψεις για να παράξει κανόνες που καλύπτουν ολόκληρο το σύνολο. Ωστόσο όπως σε κάθε κατανεμημένο σύστημα, έτσι και στο μοντέλο της Απεικόνισης/Μείωσης υπάρχει ένα κόστος για την αρχικοποίηση και το ξεκίνημα της συστοιχίας των υπολογιστών για την εκτέλεση ενός έργου. Και επειδή σε κάθε επανάληψη υπάρχουν δύο έργα Απεικόνισης/Μείωσης το συνολικό κόστος ενός έργου μπορεί να καταστεί ασύμφορο. Προς την αντιμετώπιση αυτού του ζητήματος υπάρχουν δύο κατευθύνσεις α) η μείωση των έργων Απεικόνισης/Μείωσης έτσι ώστε το πλήθος τους να μην εξαρτάται από τη φύση του συνόλου δεδομένων και β) η χρήση τεχνικών βασισμένων στη μνήμη, όπως το Spark, που μειώνουν δραματικά το κόστος ανά επανάληψη. Για την πρώτη από τις δυο κατευθύνσεις, η μείωση των επαναλήψεων σε κάποιο σταθερό αριθμό, μπορεί να επιτευχθεί με την προσωρινή αποθήκευση μέρους του χώρου αναζήτησης των κανόνων στη μνήμη, ο οποίος θα πρέπει να περιέχει μόνο εκείνους τους κανόνες που αφενός είναι ποιοτικοί, αφετέρου καλύπτουν διαφορετικό κομμάτι του συνόλου δεδομένων. Έτσι καθίσταται δυνατή η κάλυψη μεγαλύτερου μέρους του χώρου αναζήτησης με μικρότερο πλήθος επαναλήψεων. Για τη δεύτερη από τις δυο κατευθύνσεις, η χρήση τεχνικών βασισμένων στη μνήμη επιφέρει μείωση στο κόστος ανά επανάληψη, κάτι που μπορεί να οδηγήσει στη συνολική μείωση του κόστους της προσέγγισης.

Τέλος, ο αλγόριθμος αναζήτησης σχετικών ιστοσελίδων που παρουσιάστηκε στο Κεφάλαιο 4, μπορεί να χρησιμοποιηθεί για την εύρεση θεματικά σχετικών διεπαφών αναζήτησης στον Παγκόσμιο Ιστό. Στην τωρινή υλοποίησή του, ο αλγόριθμος δεδομένης μιας αρχικής ιστοσελίδας, συγκρίνει γειτονικές της ιστοσελίδες με ένα αυστηρά καθορισμένο κριτήριο ποιότητας και επιλέγει την πιο ποιοτική σε κάθε επανάληψη. Η σύγκριση γίνεται με τεχνικές ομοιότητας που βασίζονται στο κείμενο των ιστοσελίδων. Ένας πιθανός συνδυασμός της προσέγγισης αυτής με ένα έτοιμο σύνολο κανόνων παραγμένων με την προσέγγιση του Κεφάλαιο 3 μπορεί να οδηγήσει στην εύρεση θεματικά σχετικών διεπαφών αναζήτησης. Επιπρόσθετα, με την εισαγωγή των κατάλληλων ευρετικών μηχανισμών στη διαδικασία μπορεί να καταστεί δυνατή και η εύρεση όχι μόνο θεματικά αλλά δομικά όμοιων διεπαφών αναζήτησης.

Αναφορές

1. **Ward, J. S. and Barker, A.** *Undefined by data: a survey of big data definitions.* s.l. : arXiv preprint, 2013. arXiv:1309.5821.

2. **Laney, D.** *3D data management: Controlling data volume, velocity and variety.* s.l. : META Group Research Note 6, 2001.

3. *NIST Big Data Program.* [Online] http://bigdatawg.nist.gov/home.php.

4. **Bayer, M. and Laney, D.** *The importance of "big data": a definition.* s.l. : Stamford, CT: Gartner, 2012.

5. What is big data? - Bringing big data to the enterprise. [Online] http://www-01.ibm.com/software/data/bigdata/.

6. **Dijcks, J. P.** *Oracle: Big data for the enterprise.* 2012 : Oracle White Paper.

7. Intel Peer Research on Big Data Analysis. [Online] http://www.intel.com/content/www/us/en/big-data/data-insights-peer-research-report.html.

8. The Big Bang: How the Big Data Explosion Is Changing the World. [Online] http://news.microsoft.com/2013/02/11/the-big-bang-how-the-big-data-explosion-is-changing-the-world/.

9. *Dryad: distributed data-parallel programs from sequential building blocks.* **Isard, M., et al.** 3, s.l. : ACM , 2007, SIGOPS Oper Syst Rev, Vol. 41.

10. *Dryadlinq: a system for general-purpose dis- tributed data-parallel computing using a high-level language.* **Yu, Y., et al.** 2008, OSDI, Vol. 8, pp. 1-14.

11. **Dean, J. and Ghemawat, S.** MapReduce: simplified data processing on large clusters. *Communications of the ACM.* 2008, Vol. 51, 1.

12. *Clustera: an integrated computation and data management system.* **DeWitt, D., et al.** 1, 2008, Proceedings of the VLDB Endowment, Vol. 1, pp. 28-41.

13. *Hyracks: A flexible and extensible foundation for data-intensive computing.* **Borkar, V., et al.** s.l. : IEEE, 2011. Data Engineering (ICDE), 2011 IEEE 27th International Conference on. pp. 1151-1162.

14. *Pregel: a system for large-scale graph processing.* **Malewicz, G., et al.** s.l. : ACM , 2010. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. pp. 135-146.

15. Apache Spark. [Online] https://spark.apache.org/.

16. *Spark: cluster computing with working sets.* **Zaharia, M., et al.** 2010. In Proceedings of the 2nd USENIX conference on Hot topics in cloud computing.

17. *Map-reduce for machine learning on multicore.* **Chu, C., et al.** 2007. Advances in Neural Information Processing Systems 19.

18. **Zaharia, M., et al.** *Job scheduling for multi-user MapReduce clusters.* s.l. : Electrical Engineering and Computer Sciences, University of California at Berkeley, 2009. UCB/EECS-2009-55.

19. *Scale and performance in a distributed file system.* **Howard, J., et al., et al.** 1, s.l. : ACM, 1988, ACM Transactions on Computer Systems, Vol. 6, pp. 51–81.

20. *Serverless network file systems.* **Anderson, T., et al.** Colorado : s.n., 1995. In Proceedings of the 15th ACM Symposium on Operating Systems Principles. pp. 109–126.

21. *GPFS: A shared-disk file system for large computing.* **Schmuck, F. and Haskin, R.** 2002. In Proceedings of the First USENIX Conference on File and Storage Technologies. pp. 231–244.

22. *The Google File System.* **Ghemawat, S., Gobioff, H. and Leung, S.** 2003. In Proceedings of the 19th ACM Symposium on Operating Systems Principles. pp. 29-43.

23. Apache Hadoop. [Online] https://hadoop.apache.org/.

24. Apache Thrift. [Online] https://thrift.apache.org/.

25. Protocol Buffers. [Online] https://developers.google.com/protocol-buffers/.

26. Apache Avro. [Online] https://avro.apache.org/.

27. Apache Cassandra. [Online] http://cassandra.apache.org/.

28. Apache HBase. [Online] http://hbase.apache.org/.

29. mongoDB. [Online] https://www.mongodb.org/.

30. Apache CouchDB. [Online] http://couchdb.apache.org/.

31. Apache Kafka. [Online] http://kafka.apache.org/.

32. Apache Storm. [Online] https://storm.apache.org/.

33. **Furnkranz, J., Gamberger, D. and Lavrac, N.** *Foundations of rule learning.* s.l. : Springer, 2012.

34. *Generalization as search.* **Mitchell, T. M.** 2, 1982, Artificial intelligence, Vol. 18, pp. 203-226.

35. *A comparative study on methods for reducing myopia of hill-climbing search in multirelational learning.* **Castillo, L. P. and Wrobel, S.** 2004. In Proceedings of the twenty-first international conference on Machine learning. p. 19.

36. *The multi-purpose incremental learning system AQ15 and its testing application to three medical domains.* **Michalski, R. S., et al., et al.** 1986. Proc. AAAI. pp. 1-41.

37. *The CN2 induction algorithm.* **Clark, P. and Niblett, T.** 4, 1989, Machine Learning , Vol. 3, pp. 261-283.

38. *On the quest for optimal rule learning heuristics.* **Frederik, J. and Furnkranz, J.** 2010, Machine Learning, pp. 343-379.

39. *Learning efficient classification procedures and their application to chess end games.* **Quinlan, J. R.** s.l. : Springer Berlin Heidelberg, 1983, Machine learning, pp. 463-482.

40. *Roc'n rule learning - towards a better understanding of covering algorithms.* **Furnkranz, J and Flach, P.** 1, 2005, Machine Learning, Vol. 58, pp. 39-77.

41. *Rule quality measures for rule induction systems: Description and evaluation.* **Aijun, A. and Cercone, N.** 3, 2001, Computational Intelligence, Vol. 17, pp. 409-424.

42. *Discretization: An Anabling Technique.* **Liu, H., et al.** 4, 2002, Data Mining and Knowledge Discovery, Vol. 6, pp. 393-423.

43. *A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning.* **Garcia, S., et al.** 4, s.l. : IEEE, 2013, Knowledge and Data Engineering, IEEE Transactions on, Vol. 25, pp. 734-750.

44. *Supervised and Unsupervised Discretization of Continuous Features.* **Dougherty, J., Kohavi, R. and Sahami, M.** 1995. Proc. 12th Int'l Conf. Machine Learning (ICML). pp. 194-202.

45. *Building a New Taxonomy for Data Discretization Techniques.* **Bakar, A.A., Othman, Z.A. and Shuib, N.L.M.** 2009. Proc. Conf. Data Mining and Optimization (DMO). pp. 132-140.

46. **Yang, Y., Webb, G.I. and Wu, X.** Discretization Methods. *Data Mining and Knowledge Discovery Handbook.* s.l. : Springer, 2010, pp. 101-116.

47. **Han, J., Kamber, M. and Pei, J.** *Data Mining: Concepts and Techniques.* 2. s.l. : Morgan Kaufmann, 2006. Vol. The Morgan Kaufmann Series in Data Management Systems.

48. **Quinlan, J. R.** *C4.5: Programs for Machine Learning.* s.l. : Morgan Kaufmann Publishers, 1993.

49. *Multivariate Discretization by Recursive Supervised Bipartition of Graph.* **Ferrandiz, S. and Boulle, M.** 2005. Fourth Conf. Machine Learning and Data Mining. pp. 253-264.

50. *Toward Unsupervised Correlation Preserving Discretization.* **Mehta, S., Parthasarathy, S. and Yang, H.** 9, 2005, IEEE Trans. Knowledge and Data Eng, Vol. 17, pp. 1174-1185.

51. *An Entropy-Based Discretization Method for Classification Rules with Inconsistency Checking.* **Li, R. P. and Wang, Z. O.** 2002. Proc. First Int'l Conf. Machine Learning and Cybernetics. pp. 243-246.

52. *ChiMerge: Discretization of Numeric Attributes.* **Kerber, R.** 1992. Proc. Nat'l Conf. Artifical Intelligence Am. Assoc. for Artificial Intelligence. pp. 123-128.

53. *Quantization of Real Value Attributes - Rough Set and Boolean Reasoning Approach.* **Nguyen, S. H. and Skowron, A.** 1995. Second Joint Ann. Conf. Information Sciences (JCIS). pp. 34-37.

54. *A streaming parallel decision tree algorithm.* **Ben-Haim, Y. and Tom-Tov, E.** 2010, The Journal of Machine Learning Research, Vol. 11, pp. 849-872.

55. *Very Simple Classification Rules Perform Well on Most Commonly Used Datasets.* **Holte, R. C.** 1993, Machine Learning, Vol. 11, pp. 63-90.

56. **Lin, J. and Dyer, C.** *Data-intensive text processing with MapReduce.* s.l. : Morgan Claypool, 2010. pp. 1-177.

57. *On the quasi-minimal solution of the covering problem.* **Michalski, R.** Bled, Yugoslavia : s.n., 1969. In Proceedings of the 5th International Symposium on Information Processing.

58. *Upper and lower bounds on the cost of a map-reduce computation.* **Sarma, A. D., et al., et al.** 2013. In Proceedings of the VLDB Endowment. Vols. Vol 5, No 4, pp. 277-288.

59. *Multi-interval discretization of continuousvalued attributes for classification learning.* **Fayyad, U. and Irani, K.** 1993. Thirteenth International Joint Conference on Articial Intelligence. pp. 1022-1027.

60. **Bache, K. and Lichman, M.** *UCI Machine Learning Repository.* [Online] University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml.

61. *The WEKA Data Mining Software: An Update.* **Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten.** 1, 2009, SIGKDD Explorations, Vol. 11.

62. *Comparison of Machine Learning Classifiers to Statistics and Neural Networks.* **Feng, C., et al.** 1993. AI & Stats Conf. 93.

63. *A Probablistic Classification System for Predicting the Cellular Localization Sites of Proteins.* **Horton , P. and Nakai, K.** St. Louis, USA : s.n., 1996, Intelligent Systems in Molecular Biology, pp. 109-115.

64. **Evett, I. and Spiehler, E. J.** *Rule Induction in Forensic Science.* Reading, Berkshire RG7 4PN : Central Research Establishment. Home Office Forensic Science Service. .

65. *Generalized Residuals for Log-Linear Models.* **Haberman, S. J.** Boston : s.n. Proceedings of the 9th International Biometrics Conference. pp. 104-122.

66. Image Segmentation Data Set . *UCI Machine Learning Repository.* [Online] https://archive.ics.uci.edu/ml/datasets/Image+Segmentation.

67. *Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets.* **Gorman, R. P. and Sejnowski, T. J.** Neural Networks, Vol. 1, pp. 75-89.

68. **Siebert, J. P.** *Vehicle Recognition Using Rule Based Methods.* Turing Institute Research Memorandum. 1987. TIRM-87-018.

69. *Classification and Regression Trees.* **Breiman, L. and Friedman, J. H.** 1984.

70. Amazon Elastic MapReduce. [Online] http://aws.amazon.com/elasticmapreduce.

71. [Online] [Cited: March 29, 2015.] http://www.internetworldstats.com/stats.htm.

72. *Tor Project: Anonymity Online.* [Online] https://www.torproject.org/.

73. *Web-scale data integration: You can only afford to pay as you go.* **Madhavan, Jayant, S. Jeffery, Shirley Cohen, X. Dong, David Ko, Cong Yu, and Alon Halevy.** s.l. : CIDR, 2007.

74. *Crawling for domain-specific hidden web resources.* **Bergholz, André and Childlovskii, B.** s.l. : IEEE, 2003. Web Information Systems Engineering, Proceedings of the Fourth International Conference on.

75. *Structured databases on the web: Observations and implications.* **Chang, Kevin Chen-Chuan, Bin He, Chengkai Li, Mitesh Patel, and Zhen Zhang.** no 3, s.l. : ACM, 2004, SIGMOD Record, Vol. 33, pp. 61-70.

76. *Google Merchant Center.* [Online] https://www.google.com/merchants.

77. *Bing Merchant Center.* [Online] http://advertise.bingads.microsoft.com/en-us/help-topic/how-to/51085/create-a-bing-merchant-center-store.

78. *Web crawling.* **Olston, C and Najork, M.** 3, Foundations and Trends in Information Retrieval, Vol. 4, pp. 175–246.

79. *Understanding deep web search interfaces: a survey.* **Khare, Ritu, Yuan, An and Il-Yeol, Song.** 1, s.l. : ACM, 2010, ACM SIGMOD Record, Vol. 39, pp. 33-40.

80. *A hierarchical approach to model web query interfaces for web source integration.* **Dragut, E. C., et al.** 2009. Proc. VLDB Endow. Vol. 2(1).

81. *Querying capability modeling and construction of deep web sources.* **Shu, L., et al.** s.l. : Springer Berlin Heidelberg, 2007. Web Information Systems Engineering–WISE 2007. pp. 13-25.

82. *Google's deep Web crawl.* **Madhavan, J., Ko, D. and Lot, L.** 2008. n Proc. 34th Int. Conf. on Very Large Data Bases. pp. 1241–1252.

83. *A brief survey of web data extraction tools.* **Laender, A. H.F., et al.** 2, 2002, ACM Sigmod Record, Vol. 31, pp. 84-93.

84. *Roadrunner: Towards automatic data extraction from large web sites.* **Crescenzi, V., Mecca, G. and Merialdo, P.** 2001. VLDB, vol. 1. pp. 109-118.

85. *Fully automatic wrapper generation for search engines.* **Zhao, H., et al.** s.l. : ACM, 2005. Proceedings of the 14th international conference on World Wide Web. pp. 66-75.

86. *ODE: Ontology-assisted data extraction.* **Su, W., Wang, J. and Lochovsky, F. H.** 2, 2009, ACM Transactions on Database Systems, Vol. 34.

87. *Structured data on the web.* **Cafarella, Michael J., Halevy, Alon and Madhavan, Jayant.** 2011, Communications of the ACM, pp. 72-79.

88. *Query routing: Finding ways in the maze of the DeepWeb.* **Kabra, Govind, Li, Chengkai and Chang, KC-C.** 2005. Web Information Retrieval and Integration, International Workshop on Challenges in.

89. *Prequery discovery of domain-specific query forms: A survey.* **Moraes, Mauricio C, et al.** no. 8, s.l. : IEEE, 2013, Knowledge and Data Engineering, Transactions on, Vol. 25, pp. 1830-1848.

90. HTML 4.01 Specification. [Online] W3C. http://www.w3.org/TR/html401/.

91. *Learning to understand information on the Internet: An example-based approach.* **Perkowitz, M., Doorenbos, R. B., Etzioni, O., & Weld, D. S.** 2, s.l. : Springer, 1997, Journal of Intelligent Information Systems, Vol. 8, pp. 133-153.

92. *Crawling the Hidden Web.* **Raghavan, Sriram and Garcia-Molina, Hector.** 2001. 27th Int'l Conf. Very Large Data Bases (VLDB '01). pp. pp. 129-138.

93. *WISE-Cluster: Clustering E-Commerce Search Engines Automatically.* **Peng, Qian , et al.** s.l. : ACM, 2004. Sixth Ann. Int'l Workshop Web Information and Data Management. pp. 104-111.

94. *Automatically Training Form Classifiers.* **Moraes, M. C., et al.** s.l. : Springer Berlin Heidelberg, 2013. In Web Information Systems Engineering–WISE 2013. pp. 441-453.

95. *Automatic discovery of Web Query Interfaces using machine learning techniques.* **Marin-Castro, H., et al.** 1, 2013, Journal of Intelligent Information Systems, Vol. 40, pp. 85-108.

96. *Combining classifiers to identify online databases.* **Barbosa, Luciano and Freire, Juliana.** s.l. : ACM, 2007. 16th International Conference on World Wide Web. pp. 431-440.

97. *Automated discovery of search interfaces on the web.* **Cope, Jared, Craswell, Nick and Hawking, David .** s.l. : Australian Computer Society, 2003. Proceedings of the 14th Australasian database conference. Vol. 17, pp. pp. 181-189.

98. *Searching for Hidden-Web Databases.* **Barbosa, Luciano and Freire, Juliana .** Baltimore, Maryland : s.n., 2005. Eighth International Workshop on the Web and Databases.

99. *An adaptive crawler for locating hidden-web entry points.* **Barbosa, Luciano and Freire, Juliana .** s.l. : ACM, 2007. 16th international conference on World Wide Web. pp. pp. 441-450.

100. The UIUC Web integration repository. [Online] [Cited: Μάρτιος 29, 2015.] http://metaquerier.cs.uiuc.edu/repository.

101. Language Data. *Webscope.* [Online] http://webscope.sandbox.yahoo.com/catalog.php?datatype=l.

102. **Brachman, Ron.** Yahoo's WebscopeTM Data Sharing Program. [Online] http://sites.nationalacademies.org/cs/groups/depssite/documents/webpage/deps_087669.pptx..

103. Public Suffix List. [Online] Mozilla Foundation. https://publicsuffix.org/.

104. **Shuyo, Nakatani.** *Language detection library for java.* [Online] 2010. http://code. google. com/p/language-detection.

105. HTML5. *World Wide Web Consortium.* [Online] http://www.w3.org/TR/html5/.

106. Apache Hive. [Online] Apache Software Foundation. [Cited: Μάρτιος 29, 2015.] https://hive.apache.org/.

107. Usage of top level domains for websites. *Web Technology Reviews.* [Online] [Cited: Μάρτιος 29, 2015.] http://w3techs.com/technologies/overview/top_level_domain/all.

108. *Automatically Training Form Classifiers.* **Moraes, M. C., et al.** s.l. : Springer Berlin Heidelberg, 2013. Web Information Systems Engineering–WISE 2013. pp. 441-453.

109. *Swarm Intelligence.* **Beni, G. and Wang, J.** Tokyo : RSJ Press, 1989. In Proceedings Seventh Annual Meeting of the Robotics Society of Japan. pp. 425-428.

110. *The self-organizing exploratory pattern of the argentine ant.* **Deneubourg, J. L., et al.** 2, 1990, Journal of insect behavior, Vol. 3, pp. 159-168.

111. *Self-organized shortcuts in the Argentine ant.* **Goss, S., et al.** 12, 1989, Naturwissenschaften, Vol. 76, pp. 579-581.

112. **Dorigo, M. and Stutzle, T.** *Ant colony optimization.* s.l. : MIT Press, 2004.

113. *Flocks, herds, and schools: a distributed behavioral model.* **Reynolds, R.G.** 4, 1987, Computer Graphics, Vol. 21, pp. 25-34.

114. **Wilson, E. O.** *Sociobiology: the new synthesis.* s.l. : Belknap Press, 1975.

115. *Particle swarm optimization.* 1995. Proceedings of the IEEE International Joint Conference on Neural Networks. pp. 1942-1948.

116. **Baeza-Yates, R. and Ribeiro-Neto, B.** *Modern Information Retrieval.* s.l. : Addison Wesley Longman Inc, 1999.

117. **Manning, C. D., Raghavan, P. and Schütze, H.** *Introduction to information retrieval.* 2008 : Cambridge university press.

118. *Fuzzy information retrieval model revisited.* **Zadrożny, S. and Nowacka, K.** 15, 2009, Fuzzy Sets and Systems, Vol. 160, pp. 2173-2191.

119. **Landauer, T., et al., et al.** *Handbook of Latent Semantic Analysis.* s.l. : Psychology Press, 2013.

120. *Relevance weighting of search terms.* **Robertson, S.E. and Sparck Jones, K.** 3, 1976, Journal of the American Society for Information Sciences, Vol. 27, pp. 129-146.

121. *Positional language models for information retrieval.* **Lv, Y. and Zhai, C.** s.l. : ACM, 2009. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 299-306.

122. *Ant Seeker: An algorithm for enhanced web search.* **G. Kouzas, E. Kayafas, V. Loumos.** Athens : Springer , 2006. 3rd IFIP Conference on Artificial Intelligence.

123. *Computer evaluation of indexing and text processing.* **Salton, G. and Lesk, M. E.** 1, s.l. : ACM, 1968, Journal of the ACM, Vol. 15, pp. 8-36.

124. **Salton, G.** *The SMART Retrieval System - Experiments in Automatic Document Processing.* s.l. : Prentice Hall, 1971.

125. **Isaacs, J. D.** *Investigating measures for pairwise document similarity.* Hanover : Dartmouth College, 1999. CS-TR99-357.

126. *Classification of a large web page collection applying a GRNN architecture.* **Anagnostopoulos, I, et al., et al.** s.l. : Springer-Verlang, 34-41. ISCIS 03.

127. *Classifying Web Pages employing a Probabilistic Neural Network Classifier.* **Anagnostopoulos, I, et al., et al.** 2004. IEE Proceedings-Software. pp. 139-150.

128. **Foundation, Mozilla.** Public Suffix List. [Online] https://publicsuffix.org/.

129. *Phrase-based Document Similarity Based on an Index Graph Model.* **Hammouda, K. M. and Kamel, M. S.** Maebashi City, Japan : IEEE, 2002. International Conference on Data Mining (ICDM 2002).

130. *Incremental Document Clustering Using Cluster Similarity Histograms.* **Hammouda, K. M. and Kamel, M. S.** Halifax : IEEE, 2003. WIC International Conference on Web Intelligence.

131. **Dorigo, M. and Stutzle, T.** *Ant Colony Optimization.* s.l. : The MIT Press, 2004.

132. *Annotating structured data of the deep Web.* **Lu, Yiyao, et al.** s.l. : IEEE. In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. pp. 376-385.

Σύντομο Βιογραφικό

Ο Δρ. Βασίλειος Γ. Κόλιας γεννήθηκε στην Αθήνα το 1984. Αποφοίτησε από το τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων του Πανεπιστημίου Αιγαίου το 2007. Το 2015, ολοκλήρωσε τη διδακτορική του διατριβή στο Εθνικό Μετσόβιο Πολυτεχνείο με θέμα την ανάπτυξη αλγορίθμων επαγωγής κανόνων κατηγοριοποίησης για μεγάλα δεδομένα και την εφαρμογή τους στο πρόβλημα της αυτόματης αναγνώρισης διεπαφών αναζήτησης στον Παγκόσμιο Ιστό αλλά και την αναζήτηση πληροφορίας στον Παγκόσμιο Ιστό με τεχνικές εμπνευσμένες από τη φύση. Από το 2008 εργάζεται ως ερευνητικός συνεργάτης στο Εργαστήριο Τεχνολογίας Πολυμέσων της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου και συμμετέχει σε ερευνητικά προγράμματα χρηματοδοτούμενα από την Ευρωπαϊκή Ένωση. Τα ερευνητικά του ενδιαφέροντα εστιάζονται στον τομέα της μηχανικής μάθησης, της υπολογιστικής νοημοσύνης και της ανάκτησης πληροφορίας.