Αυτοματοποιημένη Κατηγοριοποίηση Δυναμικών Δεδομένων με έμφαση στις Σελίδες Διαδικτύου: μια συνδυαστική προσέγγιση

Μαρία Δ. Νιάρου

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Σχολή Επιστήμης της Πληροφορίας και Πληροφορικής

Τμήμα Αρχειονομίας, Βιβλιοθηκονομίας και Μουσειολογίας

Κέρκυρα
Δεκέμβρης, 2021

Automatic Classification Technique for Dynamic Data, focused on Webpages: a combined approach

Maria D. Niarou

DOCTORAL DISSERTATION

IONIAN UNIVERSITY

Faculty of Information Science & Informatics

Department of Archives, Library Science and Museum Studies

Corfu
December, 2021

Αυτοματοποιημένη Κατηγοριοποίηση Δυναμικών Δεδομένων με έμφαση στις Σελίδες Διαδικτύου: μια συνδυαστική προσέγγιση

Μαρία Δ. Νιάρου

Διδακτορική Διατριβή

Επιβλέπουσα:

Στάμου Σοφία, Επικ. Καθηγήτρια του Τμήματος Αρχειονομίας, Βιβλιοθηκονομίας και Μουσειολογίας του Ιονίου Πανεπιστημίου, με γνωστικό αντικείμενο «Βιβλιοθηκονομία: Πρότυπα Οργάνωσης και Συστήματα Γλωσσικής Επεξεργασίας Τεκμηρίων»

Στην **Τριμελή Συμβουλευτική Επιτροπή**, εκτός από την κύρια επιβλέπουσα, συμμετείχαν οι:

Γεργατσούλης Εμμανουήλ, Καθηγητής του Τμήματος Αρχειονομίας, Βιβλιοθηκονομίας και Μουσειολογίας του Ιονίου Πανεπιστημίου, με γνωστικό αντικείμενο «Πληροφορική με έμφαση στις Βάσεις Δεδομένων και τη Διαχείριση Μεταδεδομένων»

Παπαθεοδώρου Χρήστος, Καθηγητής του Τμήματος Ιστορίας και Φιλοσοφίας της Επιστήμης του Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών, με γνωστικό αντικείμενο «Πληροφοριακά Συστήματα Βιβλιοθηκών και Αρχείων»

Στην **Επταμελή Εξεταστική Επιτροπή**, εκτός από την Τριμελή Συμβουλευτική Επιτροπή, συμμετείχαν οι:

Βαρλάμης Ηρακλής, Αναπλ. Καθηγητής του Τμήματος Πληροφορικής και Τηλεματικής του Χαροκόπειου Πανεπιστημίου, με γνωστικό αντικείμενο «Διαχείριση Δεδομένων»

Κερμανίδου Κάτια-Λήδα, Αναπλ. Καθηγήτρια του Τμήματος Πληροφορικής του Ιονίου Πανεπιστημίου με γνωστικό αντικείμενο «Τεχνητή Νοημοσύνη με Έμφαση στη Γλωσσική Τεχνολογία»

Μαραγκουδάκης Εμμανουήλ, Καθηγητής του Τμήματος Πληροφορικής του Ιονίου Πανεπιστημίου με γνωστικό αντικείμενο «Δομές και Βάσεις Δεδομένων»

Σφακάκης Μιχάλης, Καθηγητής του Τμήματος Αρχειονομίας, Βιβλιοθηκονομίας και Μουσειολογίας του Ιονίου Πανεπιστημίου, με γνωστικό αντικείμενο «Θεωρία και Πρακτική της Θεματικής Ανάλυσης και Οργάνωσης της Πληροφορίας»

Στους κόπους μου & τους καρπούς τους!

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ ΕΙΚΟΝΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ ΠΙΝΑΚΩΝ

ΕΥΧΑΡΙΣΤΙΕΣ

Αθήνα, 2012-2021
Έκθεση, αντί ευχαριστιών!

Το κομμάτι των ευχαριστιών είναι το σημείο εκείνο κάθε δουλειάς, όπου μπορεί κάποιος να γράψει σε πρόσωπο πρώτο και με τρόπο ελεύθερο. Έτσι, αυτή τη φορά, που νιώθω να έχω πάρει το δικό μου, προσωπικό βραβείο Oscar, επιλέγω το κομμάτι αυτό να έχει μορφή έκθεσης, δηλαδή να είναι ένα ελεύθερο κείμενο, μιας καί ουσιαστικά πρόκειται για έκθεση, δηλαδή ευθεία έκφραση συναισθημάτων! Πρόκειται για τη δική μου προσωπική έκθεση, λοιπόν, και με τις δύο αυτές έννοιες! Επιδιώκοντας αυτό να είναι γνήσιο, και στην προσπάθειά μου να αποτυπώσω όσο καλύτερα γίνεται το ***πλαίσιο μέσα στο οποίο πραγματοποιείται η δουλειά*** *αυτής της διατριβής, δεν περιμένω να ολοκληρώσω την εργασία μου για να συντάξω αυτό το κείμενο. Το γράφω σχεδόν παράλληλα τα τελευταία χρόνια, σε κάθε ευκαιρία, με κάθε αφορμή, ώστε κάθε σκέψη να συνοδεύεται από την αίσθηση της στιγμής!*

Στο μυαλό μου, το περιεχόμενο της διατριβής μου είναι αποτέλεσμα δουλειάς χρόνων, είναι απόσταγμα διαδρομής που, στην πραγματικότητα, έχει αφετηρία στα πρώτα μου φοιτητικά χρόνια (2003-2008). Κατά συνέπεια, ξεκινώντας να καταγράφω την περιήγησή μου στον κόσμο της ακαδημαϊκής μελέτης[1], δεν μπορώ παρά να αναφερθώ στους ακαδημαϊκούς μου δασκάλους ***Μανόλη Γεργατσούλη*** *και* ***Χρήστο Παπαθεοδώρου*** *(με αλφαβητική σειρά), μέλη και της τριμελούς Συμβουλευτικής Επιτροπής της παρούσας διατριβής, αφού είναι από τους πρώτους που συνάντησα σε αυτή τη διαδρομή. Θα ήθελα, λοιπόν, να εκφράσω την εκτίμησή μου καταρχάς προς εσάς,* ***αγαπητοί δάσκαλοι****, για την συνεισφορά σας σε αυτή τη διατριβή, αλλά και για τα επιστημονικής, και όχι μόνο, φύσης ερεθίσματα που πήρα μέσα από τα μαθήματά σας! Με αφορμή αυτά, αντιλήφθηκα την βιβλιοθηκονομία και την αρχειονομία ως έννοιες ευρύτερες, συνώνυμες της διαχείρισης πληροφορίας, όπου και όπως αυτή συναντάται. Οι εργασίες που πραγματοποίησα στο πλαίσιο των μαθημάτων σας, διεύρυναν τα ενδιαφέροντά*

[1] *Θα κάνω κάποιες εκφραστικές αυθαιρεσίες.*

μου, γεγονός που συνέβαλε στο να φτάσω σήμερα εδώ. Και πού είναι το «εδώ»; Σίγουρα τουλάχιστον ένα βήμα πιο πέρα από αυτό που είχα στο μυαλό μου όταν ξεκίνησα τις σπουδές μου!

***Παράλληλα, ξεχωριστά και εξίσου θερμά**, θέλω να εκφραστώ και για την **επιβλέπουσα αυτής της διατριβής**, δασκάλα μου **Σοφία** (Στάμου). Για την ιστορία, χωρίς να έχουμε προηγουμένως γνωριστεί στο πλαίσιο κάποιου μαθήματος ή προσωπικά, την προσέγγισα βλέποντας πως το γνωστικό της αντικείμενο συγγενεύει με τα ενδιαφέροντά μου. Μάλιστα, κάπως ταυτίστηκα που στο βιογραφικό της έβλεπα να έχει συνδυάσει τις θεωρητικές με τις θετικές επιστήμες, καθώς ελκύουν κι εμένα εξίσου η γλώσσα και τα μαθηματικά – για να το πω με ένα απλό σχήμα. Έτσι ξεκίνησε η συνεργασία μας, που είχε σαν πρώτο αποτέλεσμα να υποστηρίξω μια διπλωματική μεταπτυχιακή εργασία άρτια και εξαιρετικά ενδιαφέρουσα! Με την ευκαιρία, θέλω να εκφράσω και εδώ τη χαρά μου για αυτή την τόσο ώριμη για εμένα δουλειά που κάναμε εκείνη την περίοδο (2010), μιας καί αυτή η δουλειά αποτέλεσε τελικώς ένα ακόμα σκαλοπάτι πριν ή για την πραγματοποίηση της παρούσας διατριβής. Όσο κι αν θέλω να αποφύγω τις «ευχαριστίες», **Σοφία, σε ευχαριστώ!** Ευχαριστώ διότι, στα μάτια μου, αυτή η διατριβή ως ακαδημαϊκό πόνημα είναι σαν καλλιτεχνική δημιουργία. Είναι αυτό που είναι, επειδή συγκεκριμένοι άνθρωποι δούλεψαν μαζί τη συγκεκριμένη χρονική περίοδο υπό τις συγκεκριμένες συνθήκες! Για χρόνια δουλεύουμε από απόσταση, και αυτό ενίοτε αποτελεί αναπόφευκτο εμπόδιο στην επικοινωνία πρακτικά, χωρίς, όμως, να αποτελεί εμπόδιο στην επικοινωνία ουσιαστικά! Σοφία, ο συμβουλευτικός σου ρόλος, η τόνωση του ηθικού, η πάντα ευδιάθετη και γεμάτη ενέργεια φωνή σου και η σιγουριά σου προς το πρόσωπό μου και τη δουλειά μας, είναι τόσο πολύτιμα όσο και η βιβλιογραφία! Η **κ. Στάμου**, στη διάρκεια αυτής της διατριβής, υπήρξε επιστημονική και ψυχολογική μου πυξίδα. Οι παρεμβάσεις της πάντα εύστοχες, καίριες και διαφωτιστικές! Η αισιοδοξία της αστείρευτη και ακλόνητη, ακόμα και στις δύσκολες στιγμές της δουλειάς και της συνεργασίας μας! Βρίσκω, ωστόσο, πως η **μεγαλύτερη συνεισφορά** της είναι η καθοδήγησή της από σκοπιά παιδαγωγική! Άλλωστε, και από ό,τι κατάλαβα, το ζητούμενο στη διάρκεια της εκπόνησης μιας διδακτορικής διατριβής - γιατί όχι και στο σχολείο, σκέφτομαι πλέον - είναι ο δάσκαλος να φωτίζει τις διασταυρώσεις που συναντά κάποιος, όχι το μονοπάτι που*

θα επιλέξει. Ζητούμενο είναι να μοιράζεται τις γνώσεις και την πείρα του, όχι να μεταδίδει στείρες πληροφορίες. Και το σημαντικότερο, να σε κάνει να νιώθεις ισάξιός του, όχι αντάξιός του. ***Σοφία, ευχαριστώ για όλα αυτά!***

Θα ήταν παράλειψη να μην αναφέρω *εδώ και την κυρία και τους κυρίους Ηρακλή Βαρλάμη, Κάτια-Λήδα Κερμανίδου, Εμμανουήλ Μαραγκουδάκη και Μιχάλη Σφακάκη (με αλφαβητική σειρά). Πρόκειται για τους ακαδημαϊκούς που ολοκληρώνουν την 7μελή Εξεταστική Επιτροπή, και τους οποίους, την ώρα που γράφω, δεν τους γνωρίζω προσωπικά· και αυτό ακριβώς το γεγονός ενισχύει τη χαρά μου που δέχονται να συμμετέχουν στην Επιτροπή, και για το χρόνο τους να μελετήσουν τη δουλειά μου.* ***Εύχομαι η παρουσίαση της διατριβής να αποδειχθεί γόνιμο έδαφος για συζήτηση ενδιαφέρουσα για όλους!***

Σε πιο προσωπικό επίπεδο, *θα ήθελα να αναφέρω τους* ***γονείς*** *μου, Δημήτρη και Αθηνά, καθώς μεγάλωσαν την αδερφή μου κι εμένα μαθαίνοντάς μας να έχουμε την εκπαίδευση και την αυτομόρφωση ψηλά στην ιεράρχηση των προτεραιοτήτων μας, και την ίδια στιγμή να είμαστε κοινωνικά ευαισθητοποιημένες, πολιτικά αφυπνισμένες και ταξικά συνειδητοποιημένες. Στη συνέχεια, μας παρείχαν τη δυνατότητα να ασχοληθούμε «επαγγελματικά» με τη μόρφωση και τη γνώση στη διάρκεια των βασικών σπουδών μας, με την έννοια ότι το πανεπιστήμιο και ό,τι συνδέεται με αυτό ήταν η «δουλειά» μας, χωρίς να χρειάζεται να μεριμνούμε οι ίδιες για τον βιοπορισμό μας. Αισθάνομαι προνομιούχα που έτυχε να έχουμε αυτήν την παροχή από τους γονείς μας, αφού ακριβώς αυτή η «επαγγελματική» ενασχόληση που μπόρεσα να έχω με τη γνώση, αυτή είναι που επιτρέπει το μετασχηματισμό της σε εργαλείο με αποτέλεσμα την βαθιά και ουσιαστική καλλιέργεια*[2]*. Παράλληλα, με τα σωστά και τα λάθη των γονιών μου, με το παράδειγμα και το αντι-παράδειγμά τους, με όσα μου προσφέρουν και όσα μου στερούν συνέβαλαν και συμβάλλουν στην εξέλιξή μου γενικώς ως ανθρώπου και ειδικώς ως μελετήτριας.* ***Μαμά***, ***μπαμπά***, *ξέρω πως κάνετε ό,τι γνωρίζετε ως καλό και ό,τι καλύτερο μπορείτε για αυτό. Εύχομαι και ελπίζω να σκέφτεστε το ίδιο για τις κόρες σας, και να μας χαίρεστε.* ***Σας αγαπώ!***

[2] *Να, και ένας από τους λόγους που στη σύγχρονη κοινωνία απαιτείται αυτό να είναι κοινωνική παροχή και όχι τυχαία ατομική συγκυρία!*

*Η **σταθερή συμμαχία** στη ζωή μου, έτσι και στη διάρκεια της διατριβής με έναν τρόπο, είναι αυτή που έχουμε με την **αδερφή** μου! Είχαμε την τύχη να κάνουμε καλή παρέα από παιδιά και, μεγαλώνοντας, επιλέγουμε συνειδητά πια να στηρίζουμε η μία την άλλη σε κάθε στάδιο, σε κάθε επίπεδο και με όποιον τρόπο μπορούμε κάθε φορά. Ακόμα κι αν έχουν συμπληρωθεί 20 χρόνια που ζούμε σε διαφορετικές πόλεις, με τα τελευταία 4 να μας χωρίζουν πάνω από δύο χιλιάδες χιλιόμετρα, πάντα χαίρεται με τη χαρά μου, λυπάται με τη λύπη μου, ενθουσιάζεται με τον ενθουσιασμό μου, απογοητεύεται με την απογοήτευσή μου. Πρακτικά, μάλλον δεν φαίνεται ιδιαιτέρως βοηθητικό αυτό, αλλά, αν το δει κάποιος από τη σκοπιά τού συμπάσχειν, είναι σπουδαίο! Κι αν τύχει εκείνη να απουσιάσει από κάποια στιγμή μου, βρίσκω στη θέση της τον γαμπρό μου[3], που συχνά στέκεται σαν αδερφός. **Ντινάκι μου, Βαγγέλη, σας κλείνω το μάτι συνωμοτικά και χαμογελώ!***

Κοντά μου έχω πάντα και τους φίλους και τις φίλες μου,** εκ των οποίων δεν είμαι βέβαιη ότι όλοι γνωρίζουν ή θυμούνται ακριβώς το θέμα αυτής της διατριβής, αλλά με στηρίζουν και με αγαπούν έτσι κι αλλιώς, με αυτό ή το άλλο θέμα, με ή χωρίς αυτήν! **Και πώς το κάνουν;** Παραμένοντας δίπλα μου όταν οι συνθήκες με κάνουν να χάνω την ισορροπία μου, κι έτσι έχω μια στέρεη βάση να πατήσω και να ισορροπήσω ξανά! Θα ήταν αδύνατο να έχει ολοκληρωθεί αυτή η δουλειά χωρίς αυτή τη **συνθήκη ζωής**! Ορισμένοι από αυτούς, συνέβαλαν και έμπρακτα στην παρούσα διατριβή, συμμετέχοντας εθελοντικά στο στάδιο της πειραματικής μελέτης, με «απουσιολόγο» τον πολυτεχνίτη Δαμιανό Βογανάτση. Κοντά σε εκείνους, θα ήθελα ονομαστικά να αναφέρω και τον Αλέκο Διαμαντούρο, γνήσιο λάτρη των ερευνητικών και όχι μόνο προκλήσεων αφού, χωρίς να είναι φίλος στενός ή συνεργάτης, χωρίς να έχει κάποιο προσωπικό όφελος, αφιέρωσε χρόνο, σκέψη και ενέργεια, και έπαιξε καθοριστικό ρόλο στην πραγματοποίηση της πειραματικής δοκιμής αυτής της δουλειάς. **Αλέκο, και στα δικά σου, με όποιον τρόπο εσύ επιλέξεις!

[3] *Είμαι σε ηλικία που με διασκεδάζουν αυτοί οι όροι δήλωσης συγγένειας!*

Κληρονομιά μου από αυτή τη διαδρομή είναι και η πρώτη δημοσίευση του πυρήνα της διατριβής, σε συνδυασμό με την εμπειρία της παρουσίασής της σε διεθνές επιστημονικό συνέδριο. Είναι ξημερώματα 1ης Απρίλη όταν λαμβάνω το e-mail της αποδοχής του άρθρου, και οι σκέψεις μου είναι δύο: πότε θα ξημερώσει για να μπορέσω να πάρω κάποιον τηλέφωνο να μοιραστώ το γεγονός χωρίς να τον ανησυχήσω, και αν υπάρχει περίπτωση να πρόκειται για πρωταπριλιάτικο αστείο! Πολύ χαρούμενη, λοιπόν, και με αυτή την εξέλιξη, μου έδωσε κουράγιο να συνεχίσω και την τέλεια αφορμή να κάνω μόνη μου ένα ταξίδι στην Πορτογαλία, την πιο κατάλληλη για εμένα στιγμή! Δούλεψα πολύ την συγκεκριμένη παρουσίαση, και το αποτέλεσμα ήταν τέτοιο, που αποτελεί τη βάση και της τελικής παρουσίασης της διατριβής. ***Ξεχωριστή εμπειρία*** *και αυτή, λοιπόν, από όλες τις πλευρές!*

Γυρνώντας εκεί όπου ξεκίνησα αυτό το κείμενο, *όσα γράφω εδώ είναι μόνο ένα μέρος όσων σκέψεων, συναισθημάτων, γεγονότων και ανθρώπων συνοδεύουν αυτή τη δουλειά. Όλα, όμως, έχουν κοινό σημείο αναφοράς:* ***τα χρόνια της διατριβής μου θα μου θυμίζουν πάντα και εκείνα της ουσιαστικής ενηλικίωσής μου.*** *Όχι γιατί το ένα έφερε το άλλο, αλλά γιατί έτυχε χρονικά να συμπίπτουν, κι έτσι μοιραία να επηρεάσει το ένα το άλλο! Άλλωστε, όταν οι υποψήφιοι διδάκτορες καλούμαστε συχνά να εργαζόμαστε παράλληλα για τα προς το ζην, πιθανώς σε άλλο αντικείμενο, σίγουρα σκληρά, εντατικά και χωρίς ουσιαστικά εργασιακά δικαιώματα, όταν η ερευνητική δουλειά (εκ-)βιάζεται από την έγνοια της επαγγελματικής πορείας και της απαιτητικής καθημερινότητας, όταν η διατριβή «κουβαλάει», χωρίς να το θέλει και χωρίς να το αξίζει, την ελπίδα για καλύτερη δουλειά, τότε δεν μπορεί παρά ωριμότερος να βρεθεί κάποιος στην έξοδο. Δεν μπορεί, παρά να ταυτιστεί με την άποψη πως η Έρευνα, ακόμα και στο πλαίσιο της διδακτορικής διατριβής, απαιτεί και είναι εργασία πλήρους απασχόλησης, και άρα πρέπει να αμείβεται ως τέτοια, όχι απλώς να ανταμείβεται ηθικά ή, στην καλύτερη περίπτωση, να εξαργυρώνεται επαγγελματικά!*

Για το τέλος, *κρατώ την υπερηφάνεια μου, με την έννοια της ικανοποίησης και της χαράς, που κατάφερα να ολοκληρώσω αυτή τη δουλειά με τον καλύτερο δυνατό για εμένα τρόπο, δεδομένων των συνθηκών! Βασικό ρόλο σε αυτό, πιστεύω έπαιξε το γεγονός ότι αποφάσισα να ξεκινήσω τη διατριβή ύστερα από ώριμη σκέψη και*

επειδή μου αρέσει να διαβάζω και να μελετάω, μου αρέσει να μαθαίνω καινούργια πράγματα και να εξελίσσω όσα ήδη γνωρίζω. Επίσης, ήξερα εξαρχής ότι πρόκειται για μακροχρόνια και επίπονη διαδικασία, ενώ είχα ήδη ξεκινήσει καί να εργάζομαι. Όμως, παρότι μπήκα σε αυτή τη διαδικασία συνειδητοποιημένη, και χωρίς να σκέφτομαι το αποτέλεσμα με την στενά ανταποδοτική έννοια, ομολογώ πως στην πορεία χρειάστηκε αρκετές φορές να επιβληθώ στον εαυτό μου προκειμένου να κρατήσω σταθερό το κέντρο μου, θυμίζοντάς μου το κίνητρό μου για να αντλώ ξανά και ξανά δύναμη και κουράγιο από αυτό. Ομολογώ πως, ιδίως μετά τον 5ο χρόνο που κάπως είχα σαν «όριο» στο μυαλό μου, χρειάστηκε αρκετή προσπάθεια για να αποδεσμεύσω την εξέλιξη της διατριβής μου από τη χρονική της διάρκεια, και την πρόοδό μου από το γραμμικό σχήμα που είχα μέχρι τότε στο μυαλό μου. Η σημερινή πραγματικότητα τριγύρω ζητάει καρπούς, αποτελέσματα εδώ και τώρα, καμιά φορά ζητάει να τα έχουμε χθες (!), και τα μετράει όλα σε αριθμούς. Πώς να μην επηρεαστώ ζώντας μέσα σε αυτή, ακόμα και αν έχω άλλη φιλοσοφία ζωής; Ανέκτησα την εστίαση και τη στόχευσή μου, όταν αποδέχτηκα πως η ερευνητική δουλειά ***είναι συνυφασμένη*** *με τη ματαίωση,* ***στηρίζεται*** *στη δοκιμή,* ***τρέφεται*** *από το λάθος,* ***απαιτεί*** *ευελιξία και, ενίοτε, επαναπροσδιορισμό. Κι όλα αυτά χρειάζονται χρόνο και τριβή.* ***Ακριβώς όπως συμβαίνει και στη ζωή!*** *Σε αυτό βοήθησαν και οι φωτεινές στιγμές, όπου ένιωθα κάτι να δημιουργείται μέσα από όσα ζυμώνονται για καιρό! Ένιωσα να εντυπωσιάζομαι η ίδια από τον εαυτό μου μέσα από αυτή τη διαδικασία! Γι' αυτό και συχνά, που κάποιοι αναρωτιούνται πώς αντέχω να το κάνω παράλληλα με ένα τόσο δεσμευτικό ωράριο δουλειάς, τους λέω πως αυτή η διατριβή είναι το βάσανο και η ψυχοθεραπεία μου μαζί!*

Βέβαια, υπάρχει και κάτι ακόμα· οι ακαδημαϊκοί δάσκαλοι λένε – μεταξύ σοβαρού και αστείου - πως, το να σκεφτείς να τα παρατήσεις, είναι ένδειξη πως έχεις εμβαθύνει σε αυτό που μελετάς (Σοφία, εσύ μου το είπες; Θυμάμαι καλά;)! Και πως, η ουσία της διατριβής ξεκινάει όταν σου λένε οι φίλοι σου να βγεις, κι εσύ προτιμάς να μελετήσεις (Μανόλη, σίγουρα εσύ μου το είπες! Θυμάμαι πολύ καλά!). Ε, λοιπόν, τα έκανα και τα δύο! Σκέφτηκα να τα παρατήσω και αρνήθηκα να βγω για να μείνω να μελετήσω μέχρι το ξημέρωμα! ***Λέτε για αυτό να τα κατάφερα;***

Από καρδιάς,
niarou.

ΠΕΡΙΛΗΨΗ

Η παρούσα διατριβή πραγματεύεται την αυτοματοποιημένη κατηγοριοποίηση δυναμικών δεδομένων, και ειδικότερα την αυτοματοποιημένη κατηγοριοποίηση σελίδων διαδικτύου μέσα από μία συνδυαστική προσέγγιση. Πρόκειται για ένα πεδίο που απασχολεί τη διεθνή ερευνητική κοινότητα από τότε που εμφανίστηκε το διαδίκτυο, καθώς βασικές πλευρές της επιστήμης των υπολογιστών, όπως είναι η διαχείριση και ανάκτηση πληροφοριών, η διαλειτουργικότητα των πηγών πληροφόρησης, αλλά και τα μοντέλα εξαγωγής πληροφοριών, μοντέλα φιλτραρίσματος περιεχομένου και αφαίρεσης διαφημίσεων, στηρίζονται στην κατηγοριοποίηση των σελίδων διαδικτύου. Τα τελευταία χρόνια, η συγκλονιστική αύξηση της απόδοσης και του χώρου μνήμης των υπολογιστών, σε συνδυασμό με την εξειδίκευση μοντέλων μηχανικής μάθησης για την ταξινόμηση κειμένων και εικόνων, αποτελούν επιπλέον λόγους για τους οποίους το ζήτημα της κατηγοριοποίησης σελίδων διαδικτύου παραμένει στο επίκεντρο του ερευνητικού ενδιαφέροντος. Ενώ, η πολυπλοκότητα που χαρακτηρίζει την αυτοματοποιημένη κατηγοριοποίηση σελίδων διαδικτύου ως διαδικασία, η ποικιλομορφία του περιεχομένου των σελίδων διαδικτύου (εικόνες διαφορετικών μεγεθών, κείμενο, υπερσύνδεσμοι κ.λπ.) και το υπολογιστικό κόστος, συνιστούν επιπρόσθετες προκλήσεις.

Κατόπιν μελέτης των προσεγγίσεων που παρουσιάζονται στη διεθνή βιβλιογραφία για τη διαχείριση του περιεχομένου του Παγκόσμιου Ιστού, διαπιστώνουμε πως οι περισσότερες από αυτές στηρίζονται κυρίως σε τεχνικές κατηγοριοποίησης κειμένων, και ορισμένες άλλες αξιοποιούν τη δομή των σελίδων. Ζητούμενο μέσα από την παρούσα διατριβή είναι να σχεδιάσουμε μια υβριδική προσέγγιση του προβλήματος της κατηγοριοποίησης σελίδων διαδικτύου, στηριζόμενοι τόσο σε κειμενικής φύσης στοιχεία όσο και σε δομικά χαρακτηριστικά. Με άλλα λόγια, η προτεινόμενη προσέγγιση στηρίζεται σε υπάρχουσες σχετικές μεθόδους, συνδυάζοντας τις τεχνικές που αξιοποιούνται στο πλαίσιό τους έτσι, ώστε οι σελίδες να κατηγοριοποιούνται ως προς το θέμα τους, αλλά και ως προς τον τύπο τους. Αυτό σημαίνει ότι η προτεινόμενη προσέγγιση αποτελεί μια συνδυαστική ενιαία μεθοδολογία κατηγοριοποίησης σελίδων διαδικτύου, η οποία στηρίζεται σε κειμενικής και δομικής

φύσης στοιχεία, γνωρίσματα και χαρακτηριστικά των υπό εξέταση σελίδων διαδικτύου.

Στόχος, μέσα από την προτεινόμενη μεθοδολογία, είναι να αντιστοιχηθεί κάθε σελίδα που εξετάζεται στην κατάλληλη κατηγορία αφότου ελεγχθούν διάφορες παράμετροι που σχετίζονται με το περιεχόμενο και τη δομή της σελίδας. Από αυτή τη σκοπιά, στο πλαίσιο της παρούσας διατριβής, σχεδιάζουμε έναν πολυδιάστατο αλγόριθμο κατηγοριοποίησης, ο οποίος αποφασίζει για τον τύπο και το θέμα κάθε σελίδας που εξετάζει. Συμπληρωματικά, παρατηρώντας ευρύτερα τις σελίδες διαδικτύου και τη δυναμική τους φύση, διευρύνουμε την «ισχύ» της προτεινόμενης μεθοδολογίας συμπεριλαμβάνοντας δύο επιπλέον αλγορίθμους, προκειμένου να παρακολουθούμε, να εντοπίζουμε και να ελέγχουμε την ανάγκη επανα-κατηγοριοποίησης των σελίδων διαδικτύου, όπου αυτό κρίνεται απαραίτητο, με βάση τις αλλαγές σε περιεχόμενο ή/και δομή που μπορεί να έχουν προκύψει. Με αυτόν τον τρόπο, καθιερώνεται ένας τακτικός έλεγχος των κατηγοριοποιημένων σελίδων, με σκοπό το αποτέλεσμα της κατηγοριοποίησης να είναι πάντα επικαιροποιημένο.

Όσον αφορά την αποτελεσματικότητα και την απόδοση της μεθοδολογίας μας, πραγματοποιούμε δοκιμαστική πειραματική αξιολόγησή της, η οποία δείχνει ότι οι σελίδες διαδικτύου κατηγοριοποιούνται ορθώς με διττό τρόπο, δηλαδή σύμφωνα με το θέμα του περιεχομένου τους και τον δομικό τους τύπο, όπως αυτός προκύπτει από τη δομή τους. Για την πληρέστερη αξιολόγηση της προτεινόμενης μεθοδολογίας, συμπληρωματικά πραγματοποιούμε συγκριτική μελέτη μεταξύ των αποτελεσμάτων του προτεινόμενου αλγορίθμου κατηγοριοποίησης και αυτών που προκύπτουν από την εφαρμογή ενός *k*-NN αλγορίθμου. Από αυτή τη συγκριτική μελέτη προκύπτει ότι η απόδοση του προτεινόμενου αλγορίθμου μπορεί να συγκριθεί και είναι αντίστοιχη αυτής ενός κλασικού αλγόριθμου κατηγοριοποίησης κειμένων.

ABSTRACT

This dissertation deals with the automated dynamic data classification, and in particular with the web pages' automated classification through a combined approach. Since the internet has appeared, this field is widely studied given that key aspects of computer science rely on web pages' classification, such as information management and retrieval, information interoperability, and also information extraction models, content filtering models and so on. Recently, the significant development of the computer performance and memory space, combined with the machine learning specialization models for text and image classification, are further reasons why the web pages' classification remains at the center of research interest. At the same time, additional challenges are the complexity of automated web pages' classification as a process, the diversity of web page content (images of different sizes, text, hyperlinks, etc.) and the cost of computing.

Looking up the World Wide Web content management approaches presented in the international literature, we find that most of them rely mainly on text categorization techniques and some others utilize the pages' structure. This dissertation aims to design a hybrid approach for the problem of automated web pages' classification, based on both textual elements and structural features. In other words, the proposed approach is built on existing works, combining techniques used in their context so that the pages are categorized in terms of their topic, but also in terms of their type. This means that the proposed approach is a combined unified web page classification methodology, which exploits web pages' textual and structural elements, features and characteristics.

To this end, the proposed methodology assigns every page examined with the appropriate category, after having checked various parameters related to the pages' content and structure. From this point of view, we design a multidimensional categorization algorithm, which decides on the type and topic of every examined page. Additionally, looking closer the web pages and their dynamic nature, we extend the potential of the proposed methodology, including two additional algorithms,

which monitor, detect and control the need to re-classify the web pages, where necessary, based on changes that may have occurred in content and/or structure. In this way, it is introduced a regular check of the classified pages, so that the classification result is always updated.

Concerning the effectiveness and the efficiency of our methodology, we carry out a pilot experimental evaluation. Through this process it is shown that web pages are correctly classified in two ways; according to the theme of their content and according to their structural type. To boost the evaluation of the proposed methodology, we additionally carry out a comparative study between the results of the proposed classification algorithm and the results obtained by the application of a *k*-NN algorithm. This comparative study shows that the performance of the proposed algorithm can be compared and is equivalent to the performance of a traditional text classification algorithm.

ΚΕΦΑΛΑΙΟ 1: Εισαγωγή

Ο Παγκόσμιος Ιστός (World Wide Web) αποτελεί πηγή δεδομένων, ο όγκος των οποίων ολοένα και αυξάνεται. Κατά συνέπεια, αυξάνονται και οι ανάγκες για αποτελεσματικότερη επεξεργασία με σκοπό την αναζήτηση και την διαχείριση των δεδομένων αυτών. Τεχνικές όπως, ο αυτόματος τρόπος ομαδοποίησης (clustering), η χρήση οντολογιών, η σημασιολογική επεξεργασία είναι αυτές που καθιστούν τον όγκο αυτό διαχειρίσιμο και παράλληλα αυτές που αποτελούν ένα από τα τρέχοντα αντικείμενα μελέτης στο χώρο της ανάκτησης πληροφοριών από το Διαδίκτυο. Ακόμα μεγαλύτερη πρόκληση στο χώρο της διαχείρισης και ανάκτησης πληροφοριών αποτελεί η κατηγοριοποίηση δυναμικών διαδικτυακών δεδομένων, αφού η φύση τους και μόνο δημιουργεί νέες απαιτήσεις. Ειδικότερα, πρόκειται για δεδομένα δυναμικά, δηλαδή δεδομένα που χαρακτηρίζονται από μεταβολή του όγκου τους και του περιεχομένου τους, με απρόβλεπτο τρόπο, σε απρόβλεπτο βαθμό και με απρόβλεπτο ρυθμό. Αξιοσημείωτο είναι επίσης το γεγονός ότι τα δεδομένα αυτά χαρακτηρίζονται από πληθώρα μορφών˙ μπορούν να είναι δεδομένα κειμένου, δεδομένα εικόνας ή/και ήχου, δεδομένα χρηστών κοινωνικών δικτύων κ.ο.κ. Κατά συνέπεια, οι παραδοσιακές προσεγγίσεις για αυτοματοποιημένο τρόπο διαχείρισης και κατηγοριοποίησης δεδομένων αδυνατούν να ικανοποιήσουν τις ιδιαίτερες απαιτήσεις των δυναμικών δεδομένων καθώς οι πρώτες βασίζονται κυρίως στη στατιστική ανάλυση και επεξεργασία των τελευταίων.

Οι υπάρχουσες τεχνικές για την κατηγοριοποίηση των σελίδων διαδικτύου, στην πλειονότητά τους, αξιοποιούν είτε το περιεχόμενο των σελίδων είτε τη δομή τους, και πολλές από αυτές έχουν αποδειχθεί αποτελεσματικές. Ωστόσο, καθώς οι απαιτήσεις αυξάνονται, οι παραδοσιακές τεχνικές εμφανίζονται ανεπαρκείς, είτε γιατί απαιτούν χρόνο για την «εκπαίδευση» του κατηγοριοποιητή, είτε γιατί συχνά στηρίζονται σε στοιχεία σχετικά αλλά με πλεονάζουσα-περιττή πληροφορία. Η μεθοδολογία που προτείνουμε στην παρούσα διδακτορική διατριβή αντιμετωπίζει την αδυναμία αυτή. Προτείνουμε έναν αυτοματοποιημένο τρόπο διαχείρισης και κατηγοριοποίησης των σελίδων διαδικτύου μέσω μιας συνδυαστικής προσέγγισης, αυτής της δομικής και σημασιολογικής ανάλυσης και επεξεργασίας των δεδομένων.

Με άλλα λόγια, προτείνουμε μια ολοκληρωμένη μεθοδολογία για την κατηγοριοποίηση των σελίδων διαδικτύου, με τρόπο αυτοματοποιημένο και χωρίς να απαιτείται φάση «εκπαίδευσης» του κατηγοριοποιητή, η οποία καταφέρνει να χαρακτηρίσει κάθε σελίδα υπό επεξεργασία ως προς το θέμα της και ως προς τον τύπο της. Συμπληρωματικά, λαμβάνοντας υπόψη τη δυναμική φύση των δεδομένων αυτών, η προτεινόμενη μεθοδολογία καταφέρνει να εντοπίσει ενδεχόμενες αλλαγές στη δομή ή/και στο περιεχόμενο των σελίδων, με αποτέλεσμα η κατηγοριοποίηση των σελίδων να παραμένει επικαιροποιημένη.

Αντικείμενο Διατριβής

Στόχος της παρούσας διατριβής είναι ο σχεδιασμός μιας πρότυπης μεθοδολογίας κατηγοριοποίησης δυναμικών δεδομένων, η οποία συνδυαστικά αξιοποιεί στοιχεία κειμενικής φύσης και γνωρίσματα δομικού χαρακτήρα, προκειμένου να κατηγοριοποιήσει ένα σύνολο δυναμικών δεδομένων. Μάλιστα, η προτεινόμενη μεθοδολογία είναι αποδεσμευμένη από τη φάση της «εκπαίδευσης» ενός κατηγοριοποιητή με δείγμα δεδομένων. Αυτό σημαίνει ότι μπορεί να κατηγοριοποιήσει ιστοσελίδες σε πραγματικό χρόνο. Τα δεδομένα που απασχολούν την παρούσα διατριβή είναι οι σελίδες διαδικτύου, καθώς πρόκειται για δεδομένα που είναι ελεύθερα διαθέσιμα σε όλους. Έτσι, μπορούν να αξιοποιηθούν και να υποστούν επεξεργασία χωρίς περιορισμούς. Η συνεισφορά της προτεινόμενης μεθοδολογίας επεκτείνεται μέσα από το γεγονός ότι μπορεί, με τις απαραίτητες παραμετροποιήσεις, να εφαρμοστεί και σε συλλογές άλλου τύπου δυναμικών δεδομένων, όχι απαραίτητα διαδικτυακών.

Αναλυτικότερα, στο πλαίσιο της παρούσας διατριβής μελετάμε τις τεχνικές κατηγοριοποίησης σελίδων διαδικτύου, αρχικώς κάνοντας μια βιβλιογραφική επισκόπηση του θέματος. Μέσα από αυτή τη διαδικασία, επιδιώκουμε να συνοψίσουμε τα δεδομένα μας, να ορίσουμε ποιο είναι το πρόβλημα και ποιο είναι το ζητούμενο. Παράλληλα, μελετώντας τις τεχνικές που έχουν προταθεί και αξιοποιηθεί μέχρι τώρα, εντοπίζουμε ποια είναι τα κενά και ποιες οι αδυναμίες γύρω από αυτό το ζήτημα. Στη συνέχεια, σχεδιάζουμε την προτεινόμενη μεθοδολογία,

στηριζόμενοι στα αποτελέσματα της διεθνούς ερευνητικής δουλειάς και επιδιώκοντας η μεθοδολογία να ανταποκρίνεται στις νέες απαιτήσεις. Η προτεινόμενη μεθοδολογία στηρίζεται στη χρήση γνωστών εργαλείων και τεχνικών, οι οποίες συνδυάζονται με πρωτότυπο τρόπο, και ολοκληρώνεται μέσα από τρεις ανεξάρτητους αλγορίθμους που λειτουργούν συμπληρωματικά. Ο πρώτος, κατά σειρά σχεδίασης και παρουσίασης, αλγόριθμος καθιστά δυνατή την πολυδιάστατη κατηγοριοποίηση των σελίδων διαδικτύου, όπου με τον όρο «πολυδιάστατη» εννοείται ότι στηρίζεται σε περισσότερες από μία παραμέτρους. Έτσι, πρόκειται για μια τεχνική που με τρόπο αυτοματοποιημένο κατηγοριοποιεί δομικά και θεματικά τις σελίδες διαδικτύου, και μάλιστα χωρίς να απαιτείται φάση «εκπαίδευσης». Η κατηγοριοποίηση επιτυγχάνεται μέσα από δύο ξεχωριστές αλλά συμπληρωματικές διαδικασίες, όπου αξιοποιούνται στοιχεία και γνωρίσματα που εντοπίζονται και εξάγονται εύκολα και είναι καθολικά. Ο δεύτερος αλγόριθμος ελέγχει τη δυναμικότητα των σελίδων που έχουν ήδη κατηγοριοποιηθεί, και εντοπίζει εκείνες που χρειάζονται επανα-κατηγοριοποίηση μετά από ένα ορισμένο χρονικό διάστημα. Για τον σκοπό αυτό, αρχικώς εντοπίζει τις σελίδες που έχουν αλλάξει, ύστερα υπολογίζει το βαθμό αλλαγής τους και τελικώς «στέλνει» για επανα-κατηγοριοποίηση εκείνες που χρειάζεται να εξεταστούν ξανά. Αφετηρία αποτελεί το γεγονός ότι το περιεχόμενο του διαδικτύου είναι δυναμικό, και άρα παρουσιάζονται ποικίλες αλλαγές καί στις σελίδες διαδικτύου είτε ως προς τη δομή είτε ως προς το περιεχόμενό τους, με αποτέλεσμα η επικαιροποίηση της κατηγοριοποίησης να αποτελεί ένα επιπλέον ζητούμενο κατά τον σχεδιασμό σύγχρονων τεχνικών. Τέλος, ο τρίτος αλγόριθμος σχεδιάζεται για να βελτιώσει τη λειτουργία του δεύτερου αλγορίθμου, αφού υπολογίζοντας το ρυθμό μεταβολής των σελίδων, μας επιτρέπει το βέλτιστο σχεδιασμό της διαδικασίας επανα-κατηγοριοποίησης. Αυτό έχει σαν αποτέλεσμα την εξοικονόμηση χρόνου και υπολογιστικών πόρων.

Εκτός από το σχεδιασμό της προτεινόμενης μεθοδολογίας, στο πλαίσιο της παρούσας διατριβής πραγματοποιούμε και την πειραματική δοκιμή της, με σκοπό τον έλεγχο της αποτελεσματικότητάς της. Συμπληρωματικά, στο πλαίσιο της πειραματικής μελέτης της μεθοδολογίας μας, πραγματοποιούμε συγκριτική μελέτη των αποτελεσμάτων που παίρνουμε από αυτήν με αυτά που δίνει η εφαρμογή ενός

παραδοσιακού αλγορίθμου κατηγοριοποίησης στο ίδιο σύνολο πειραματικών δεδομένων. Η πειραματική δοκιμή γίνεται σε ένα μικρής κλίμακας (για τα δεδομένα του διαδικτύου) σύνολο δεδομένων, αξιοποιώντας εργαλεία που διατίθενται ελεύθερα στο διαδίκτυο, και αποδεικνύει ότι τα βήματα της προτεινόμενης μεθοδολογίας οδηγούν στην κατηγοριοποίηση των σελίδων, και μάλιστα αποδίδοντας αυτές στη σωστή κατηγορία.

Συνεισφορά Διατριβής

Ύστερα από επισκόπηση της διεθνούς βιβλιογραφίας σχετικά με τη φύση των διαδικτυακών δεδομένων και την κατηγοριοποίησή τους, μέσα από την παρούσα διατριβή προτείνουμε τη συνδυαστική κατηγοριοποίησή τους αξιοποιώντας κειμενικά στοιχεία τους και δομικά χαρακτηριστικά γνωρίσματα. Λαμβάνοντας παράλληλα υπόψη τη δυναμικότητα των δεδομένων αυτών, και στοχεύοντας στην επικαιροποίηση της κατηγοριοποίησής τους σε βάθος χρόνου, προτείνουμε συμπληρωματικά μία τεχνική για την επανα-κατηγοριοποίησή τους, όταν και όπου αυτό είναι απαραίτητο, και μία για τον εντοπισμό εκείνων των σελίδων, οι οποίες χρειάζεται να ελέγχονται για ενδεχόμενες αλλαγές ανά διαστήματα. Ο καρπός όσων μελετήθηκαν και παρουσιάζονται σε αυτή τη δουλειά, συνοψίζονται ως εξής:

- Η συνδυαστική αξιοποίηση δομικών χαρακτηριστικών γνωρισμάτων και κειμενικών στοιχείων, μπορεί να οδηγήσει σε πληρέστερη κατηγοριοποίηση των σελίδων διαδικτύου, γεγονός που συμβάλλει σημαντικά στη διαδικασία της ανάκτησής τους.
- Η αξιοποίηση απλών, εύκολα προσβάσιμων και ενιαίων χαρακτηριστικών ή/και στοιχείων των σελίδων διαδικτύου, μπορούν να επιταχύνουν τη διαδικασία κατηγοριοποίησής τους.
- Η δυναμικότητα των σελίδων διαδικτύου είναι εγγενές χαρακτηριστικό τους που πρέπει να λαμβάνεται υπόψη κατά το σχεδιασμό σύγχρονων τεχνικών διαχείρισης και οργάνωσής τους, προκειμένου το αποτέλεσμα να είναι πάντα επικαιροποιημένο.
- Ακόμα και μέσω απλών τεχνικών ελέγχου και υπολογισμού ομοιότητας μπορούν να επιτευχθούν ο έλεγχος και ο υπολογισμός του βαθμού αλλαγής

δυναμικών δεδομένων, και να οδηγήσουν στην επικαιροποίηση της κατηγοριοποίησής τους.

- Ο υπολογισμός του ρυθμού αλλαγής των δεδομένων, μπορεί να βελτιώσει τη διαδικασία επανα-κατηγοριοποίσής τους, όπου αυτό κρίνεται απαραίτητο, καθώς επιτρέπει τον καθορισμό της βέλτιστης πολιτικής επανα-κατηγοριοποίησης.
- Οι σύγχρονες τεχνικές διαχείρισης και οργάνωσης των σελίδων διαδικτύου, μπορούν να γίνουν πιο αποτελεσματικές και συμφέρουσες από πλευράς υπολογιστικών πόρων και ενέργειας όταν στηρίζονται σε διαδικασίες που δεν απαιτούν την αποθήκευσή τους.

Εκτός από τα παραπάνω, η συνεισφορά της διατριβής εντοπίζεται στο σχεδιασμό και την υλοποίηση της προτεινόμενης αλγοριθμικής τεχνικής για την κατηγοριοποίηση των διαδικτυακών σελίδων. Οι τεχνικές αυτές, στηρίζονται στη συνδυαστική αξιοποίηση της δομής και του περιεχομένου των σελίδων, προκειμένου να κατηγοριοποιηθούν με τρόπο αυτοματοποιημένο και χωρίς να απαιτείται η αποθήκευσή τους. Επιπρόσθετα, προτείνεται τεχνική για την εκτίμηση ενδεχόμενης ανάγκης επανα-κατηγοριοποίησης της κάθε σελίδας ξεχωριστά μετά από ένα ορισμένο χρονικό διάστημα και δεδομένης της δυναμικής φύσης των δεδομένων διαδικτύου, που αλλάζουν με απρόβλεπτο τρόπο. Επίσης, προτείνεται τεχνική διαχείρισης του απρόβλεπτου, καί από πλευράς ρυθμού, χαρακτήρα που διέπει τη δυναμική φύση των δεδομένων διαδικτύου. Επιστέγασμα όλων των παραπάνω αποτελεί η πειραματική αξιολόγηση της προτεινόμενης μεθοδολογίας σε ένα δείγμα δεδομένων μικρής κλίμακας, που επιβεβαιώνει τη λειτουργικότητά της.

Τέλος, ο πυρήνας της διατριβής παρουσιάστηκε στο 18th International Conference on Building and Exploring Web Based Environments (WEB 2020), υπό τον τίτλο «*A Combined Approach to Dynamic Web Page Classification: Merging Structure and Content*» [I], όπου και διακρίθηκε ως Best Conference Paper.

Δομή Διατριβής

Η παρούσα διατριβή αποτελείται από 6 Κεφάλαια. Στο παρόν Κεφάλαιο (**ΚΕΦΑΛΑΙΟ 1**: Εισαγωγή) γίνεται σύντομη παρουσίαση του αντικειμένου μελέτης, των στόχων και της συνεισφοράς της διατριβής.

Στο **ΚΕΦΑΛΑΙΟ 2**: Χαρακτηριστικά και Ιδιαιτερότητες Διαδικτύου, γίνεται σύντομη ιστορική αναδρομή στην εξέλιξη του παγκόσμιου ιστού, από το *WEB-of-Data* στο *WEB-of-Things*. Στη συνέχεια, γίνεται αναφορά στα βασικά χαρακτηριστικά του διαδικτύου, όπως είναι η διασυνδεσιμότητα των δεδομένων του, η δυναμικότητά τους, η ποικιλομορφία και η ετερογένεια που τα διακρίνουν, αλλά και η αμφίβολη, συχνά, ποιότητά τους. Το Κεφάλαιο ολοκληρώνεται με την παρουσίαση των σύγχρονων προκλήσεων, αναφορικά με τη διαχείριση των δεδομένων του παγκόσμιου ιστού. Η παρουσίασή τους γίνεται σε συνάρτηση με τις δυσκολίες που πρέπει να αντιμετωπιστούν, δεδομένων των ιδιαίτερων χαρακτηριστικών τους και προκειμένου να είναι εύκολα και αποτελεσματικά διαχειρίσιμα.

Στο **ΚΕΦΑΛΑΙΟ 3**: Τεχνικές Κατηγοριοποίησης Διαδικτυακών Δεδομένων, γίνεται αναφορά στις ήδη υπάρχουσες τεχνικές για την κατηγοριοποίηση των δεδομένων διαδικτύου. Αρχικώς, δίνεται ορισμός του όρου *κατηγοριοποίηση* και γίνεται αναφορά στη χρησιμότητά της και πώς συνδέεται με άλλα επιστημονικά πεδία. Στη συνέχεια, παρουσιάζονται ορισμένοι από τους πιο διαδεδομένους αλγορίθμους αυτόματης κατηγοριοποίησης κειμένων, στους οποίους βασίζονται με τη σειρά τους οι παραδοσιακές τεχνικές κατηγοριοποίησης σελίδων διαδικτύου, καθώς και χαρακτηριστικές τεχνικές κατηγοριοποίησης σελίδων διαδικτύου βάσει κειμενικής πληροφορίας και βάσει δομικών χαρακτηριστικών. Σκοπός του Κεφαλαίου είναι να οριστεί το πλαίσιο μέσα στο οποίο κινείται η έρευνά μας, καθώς και οι βασικές τεχνικές που αποτελούν αφετηρία και βάση για το σχεδιασμό της προτεινόμενης μεθοδολογίας.

Το **ΚΕΦΑΛΑΙΟ 4**: Προτεινόμενη Μεθοδολογία, αποτελεί τον πυρήνα της παρούσας εργασίας. Σε αυτό, γίνεται η αναλυτική παρουσίαση της προτεινόμενης μεθοδολογίας, ενώ παράλληλα τεκμηριώνονται μία-μία οι επιλογές που έγιναν για

το σχεδιασμό της, καθώς και οι διαδικασίες που η ίδια περιλαμβάνει. Συγκεκριμένα, περιγράφονται και τεκμηριώνονται οι διαδικασίες και τα βήματα των τριών ξεχωριστών αλλά συμπληρωτικών αλγορίθμων που απαρτίζουν τη μεθοδολογία, καθένας εκ των οποίων συνιστά ολοκληρωμένη και αυτοτελή διαδικασία. Σκοπός της προτεινόμενης μεθοδολογίας είναι η αυτοματοποιημένη κατηγοριοποίηση των σελίδων διαδικτύου ως προς τον τύπο και το θέμα τους, μέσω της δομικής και σημασιολογικής τους ανάλυσης και επεξεργασίας. Συμπληρωματικά, η μεθοδολογία που προτείνεται αντιμετωπίζει τη δυναμικότητα των δεδομένων αυτών, αξιολογώντας την ανάγκη για επανα-κατηγοριοποίησή τους και εκτιμώντας τη συχνότητα που χρειάζεται να γίνεται ο σχετικός έλεγχος.

Στο **ΚΕΦΑΛΑΙΟ 5**: Πειραματική Αξιολόγηση, περιγράφονται η πειραματική δοκιμή και μελέτη της προτεινόμενης μεθοδολογίας, καθώς και τα αποτελέσματα που ελήφθησαν από αυτήν. Για την πραγματοποίηση του δοκιμαστικού ελέγχου, η προτεινόμενη μεθοδολογία εφαρμόζεται σε ένα μικρής κλίμακας σύνολο δεδομένων. Επιπλέον, πραγματοποιείται και συγκριτική μελέτη των αποτελεσμάτων που καταγράφονται από την εφαρμογή της με αυτά που δίνει η εφαρμογή ενός παραδοσιακού αλγορίθμου κατηγοριοποίησης στο ίδιο σύνολο πειραματικών δεδομένων. Σκοπός της πειραματικής δοκιμής είναι να αποδειχθεί η ορθότητα της προτεινόμενης μεθοδολογίας, δηλαδή ότι λειτουργεί, όπως και η απόδοσή της, δηλαδή ότι λειτουργεί σωστά. Στο εν λόγω Κεφάλαιο, περιγράφονται αναλυτικά η διαδικασία συλλογής των πειραματικών δεδομένων, πώς εφαρμόστηκε η μεθοδολογία σε αυτά, καθώς και τα αποτελέσματα από την εφαρμογή και τη συγκριτική μελέτη. Σε ξεχωριστή Ενότητα του Κεφαλαίου, για την ολοκληρωμένη αξιολόγηση της μεθοδολογίας, περιγράφεται ο ποιοτικός έλεγχος των αποτελεσμάτων της κατηγοριοποίησης βάσει μετρικών αξιολόγησης ποιότητας.

Στο

ΚΕΦΑΛΑΙΟ 6: Συμπεράσματα, γίνεται η αποτίμηση του έργου, παρουσιάζονται τα συμπεράσματα της ερευνητικής δουλειάς, καθώς και ορισμένες από τις μελλοντικές κατευθύνσεις σε συνέχεια της παρούσας ερευνητικής μελέτης.

ΚΕΦΑΛΑΙΟ 2: Χαρακτηριστικά και Ιδιαιτερότητες Διαδικτύου

2.1. Εισαγωγή

Στις Ενότητες αυτού του Κεφαλαίου, γίνεται σύντομη ιστορική αναδρομή στην εξέλιξη του παγκόσμιου ιστού, καθώς και αναφορά στα βασικά χαρακτηριστικά του, όπως είναι η διασυνδεσιμότητα των δεδομένων, η δυναμικότητά τους, η ποικιλομορφία και η ετερογένεια που τα διακρίνουν, καθώς και στην ποιότητά τους που συχνά μπορεί να είναι αμφίβολη. Το Κεφάλαιο ολοκληρώνεται με την παρουσίαση των σύγχρονων προκλήσεων, αναφορικά με τη διαχείριση των δεδομένων του παγκόσμιου ιστού, όπως αυτές προκύπτουν από τις δυσκολίες που προκαλεί η ιδιαίτερη φύση τους.

Στο σημείο αυτό, είναι χρήσιμο να διευκρινιστεί πως με τον όρο «***Παγκόσμιος Ιστός***» ορίζεται ο τρόπος με τον οποίο είναι οργανωμένες οι πληροφορίες, ενώ με τον όρο «***Διαδίκτυο***» ορίζεται η φυσική υποδομή για την ανάπτυξη του Παγκόσμιου Ιστού. Βέβαια, καθώς οι λειτουργίες του Διαδικτύου, σε μεγάλο βαθμό, εκτελούνται μέσα από τον Παγκόσμιο Ιστό, εύκολα εξηγείται και το γεγονός ότι αυτοί οι δύο όροι ταυτίζονται εννοιολογικά στην αντίληψη των ανθρώπων. Έτσι, και στην παρούσα εργασία, οι όροι «Διαδίκτυο» και «Παγκόσμιος Ιστός» χρησιμοποιούνται ως όροι εννοιολογικά ταυτόσημοι και με την έννοια των οργανωμένων πληροφοριών.

2.2. Ιστορική Αναδρομή

Η ιστορία του διαδικτύου (Internet) ξεκινάει από την ανάπτυξη των υπολογιστών, και η δημιουργία του προκύπτει από την ανάγκη να είναι δυνατή η ανταλλαγή δεδομένων μεταξύ δύο ή περισσότερων υπολογιστών. Ωστόσο, η δυνατότητα ανταλλαγής δεδομένων μεταξύ υπολογιστών που μπορεί να βρίσκονται οπουδήποτε στον κόσμο δόθηκε το 1990, οπότε ο ερευνητής Tim Berners-Lee επινόησε το πρωτόκολλο ***HTTP*** (**H**yper**T**ext **T**ransfer **P**rotocol), Πρωτόκολλο Μεταφοράς Υπερκειμένου, και τη γλώσσα ***HTML*** (**H**yper**T**ext **M**arkup **L**anguage), Γλώσσα

Σήμανσης Υπερκειμένου, προκειμένου να μπορεί να διαμοιράζεται εύκολα και γρήγορα τα αποτελέσματα των πειραμάτων του [12]. Σκοπός της δημιουργίας τού Πρωτοκόλλου αυτού είναι να μπορέσουν οι *φυλλομετρητές* (browsers) τού Παγκόσμιου Ιστού να μεταφέρουν δεδομένα ανάμεσα σε έναν *διακομιστή* (server) κι έναν *πελάτη* (client), ενώ η HTML αποτελεί την κύρια γλώσσα σήμανσης για τις ιστοσελίδες, μέσα από τα *στοιχεία* (elements) της οποίας ορίζονται τα βασικά δομικά στοιχεία των σελίδων.

Από την παρθενική εμφάνιση του διαδικτύου μέχρι σήμερα, η εξέλιξή του είναι ραγδαία. Η πρώτη έκδοσή του, ***Web1.0***, γνωστή ως *Web-of-Data*, δηλαδή Παγκόσμιος Ιστός Δεδομένων, αφορά αποκλειστικά τον διαμοιρασμό δεδομένων. Σε αυτό το διαδικτυακό περιβάλλον, το κύριο χαρακτηριστικό είναι πως ο χρήστης μπορεί απλώς να έχει πρόσβαση σε αυτά από τον προσωπικό του υπολογιστή. Σε εκείνο το στάδιο, το διαδίκτυο εμφανίζεται σαν ένα «τέχνεργο» που μπαίνει στη ζωή των ανθρώπων. Η χρήση του είναι δυνατή από κάθε υπολογιστή βασικών τεχνικών χαρακτηριστικών, αλλά οι χρήστες του δεν έχουν κάποια συμμετοχή. Η μηχανή αναζήτησης είναι το μέσο εκείνο που διασφαλίζει την ανάκτηση των πληροφοριών, ώστε αυτές να μπορούν να αναγνωστούν από το χρήστη, *Read-only-Web* [35].

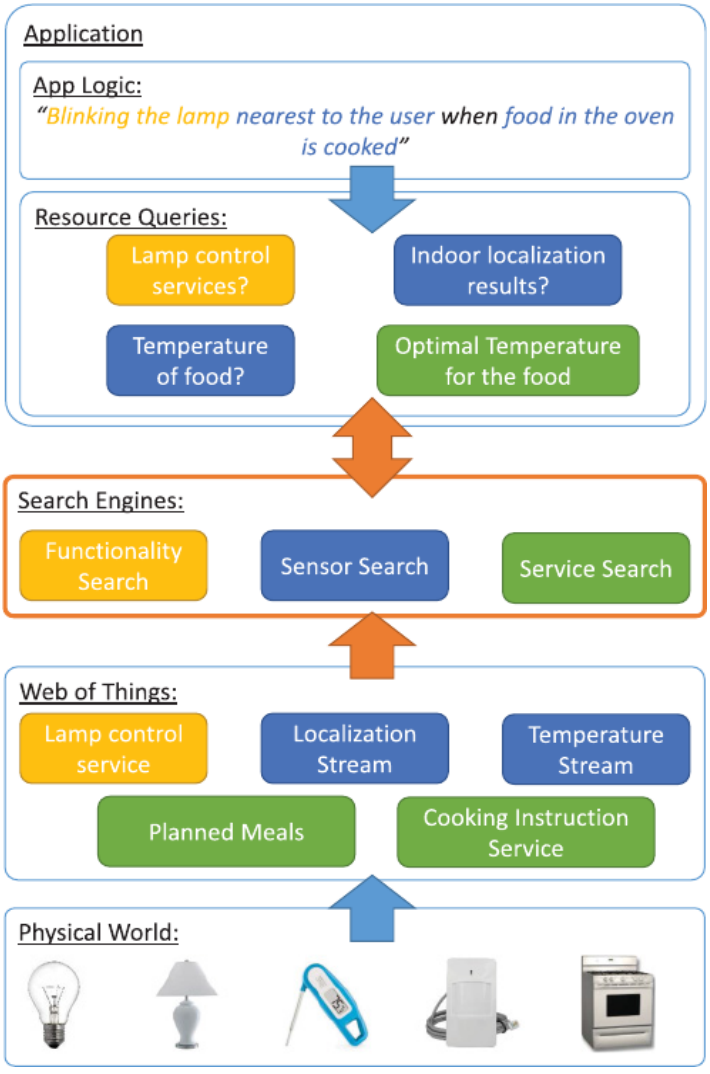
Στο τέλος της δεκαετίας του 1990, ο παγκόσμιος ιστός αρχίζει να παίρνει άλλη μορφή, ***Web2.0***, με την οποία δίνεται η δυνατότητα διάδρασης στους χρήστες και διευρύνεται ο τρόπος που μπορούν οι άνθρωποι να αξιοποιήσουν την τεχνολογία. Ωστόσο, δεν έχει ως τώρα δοθεί κάποιος σαφής ορισμός για εκείνη τη φάση εξέλιξης του παγκόσμιου ιστού [53]. Σημαντικό όφελος από τη χρήση τού Web2.0, είναι η ευκολία που χαρακτηρίζει τη χρήση του, καθώς και η χρήση εφαρμογών χωρίς να είναι απαραίτητη η εγκατάστασή τους στον υπολογιστή. Χαρακτηριστικό του Web2.0 είναι επίσης, η εμφάνιση των *wikis*. Πρόκειται για ένα λογισμικό, που επιτρέπει στους χρήστες που έχουν πρόσβαση σε αυτό να δημιουργούν περιεχόμενο συνεργατικά, να το επεξεργάζονται και να το κατηγοριοποιούν σε πραγματικό χρόνο, κάνοντας αναφορά σε ποικίλο υλικό [57].

Επόμενη εξελικτικά τομή στην ιστορία του παγκόσμιου ιστού είναι η περίοδος όπου γίνεται γνωστός και ως ***σημασιολογικός ιστός***, *semantic web*, μέσω του οποίου δίνεται η δυνατότητα στους υπολογιστές να «καταλαβαίνουν» τι «εννοεί» ο χρήστης. Το γεγονός αυτό δημιουργεί νέες για την εποχή δυνατότητες, αφού πλέον οι υπολογιστές, μέσω των *πρακτόρων* (*agents*), δεν επιστρέφουν «τυφλά» αποτελέσματα σύμφωνα με το λεκτικώς εκφρασμένο πληροφοριακό αίτημα του χρήστη, αλλά μπορούν εύκολα να εκτελούν εξελιγμένες εργασίες για τους χρήστες. Απαραίτητη προϋπόθεση για τις λειτουργίες του σημασιολογικού ιστού είναι η πρόσβαση των ηλεκτρονικών υπολογιστών σε δομημένες συλλογές πληροφοριών και σύνολα κανόνων για την εξαγωγή συμπερασμάτων, που να μπορούν να χρησιμοποιηθούν για τη διεξαγωγή αυτοματοποιημένης συλλογιστικής. [15].

Σε συνέχεια του σημασιολογικού ιστού έρχεται ο ***κοινωνικός ιστός***, *social web*. Σε αυτή τη φάση εξέλιξης του παγκόσμιου ιστού, εκείνο που χαρακτηρίζει τους ιστοτόπους είναι πως, μέσω αυτών, δίνεται η δυνατότητα στους χρήστες να σχηματίσουν *διαδικτυακές κοινότητες*, *online communities,* και να «μοιραστούν» περιεχόμενο δημιουργημένο από τους ίδιους, όπως φωτογραφίες, σελιδοδείκτες, πολυμεσικό υλικό, δραστηριότητες, κείμενο κ.α [43]. Στο σημείο αυτό, θα ήταν παράλειψη να μην αναφέρουμε πως, με την πάροδο του χρόνου, εκτός από τη θετική πλευρά αυτής εξέλιξης, έρχεται και η αρνητική, καθώς πολλοί χρήστες αφιερώνουν ιδιαιτέρως πολύ χρόνο στις σελίδες των κοινωνικών δικτύων, «χτίζοντας» την διαδικτυακή τους εικόνα και δημοσιεύοντας πολλές και λεπτομερείς πληροφορίες για τη ζωή τους, με ό,τι αυτό συνεπάγεται. Παράλληλα, φαίνεται να χάνεται το όριο για το τι είναι κατάλληλο και τι όχι για δημοσίευση, γεγονός που συνιστά πρόσφορο έδαφος για την εμφάνιση και ανάπτυξη του ηλεκτρονικού εγκλήματος. Βεβαίως, αυτή η πλευρά της εξέλιξης του διαδικτύου δεν αποτελεί μέρος του αντικειμένου της παρούσας διατριβής, για αυτό και δεν γίνεται εκτενής αναφορά.

Η σύγχρονη εκδοχή του παγκόσμιου ιστού, γνωστή ως ***web of things***, *ιστός των πραγμάτων,* χαρακτηρίζεται από το γεγονός ότι οι μηχανές αναζήτησης επιτρέπουν την αναζήτηση και τον εντοπισμό πραγματικών οντοτήτων με συγκεκριμένες ιδιότητες [58]. Παράλληλα, ολοένα και περισσότερες «έξυπνες» συσκευές κάθε

είδους διαλειτουργούν. Με άλλα λόγια, επικοινωνούν μεταξύ τους και «μοιράζονται» δεδομένα μέσω του παγκόσμιου ιστού στον ίδιο χρόνο [88]. Στην πράξη, η εξέλιξη αυτή του παγκόσμιου ιστού σε συνδυασμό με την τεχνολογική πρόοδο, επιτρέπει σε ολοένα και περισσότερα φυσικά αντικείμενα να είναι συνδεδεμένα στο διαδίκτυο και να παρέχουν τις υπηρεσίες τους μέσω αυτού. Για την πλήρη αξιοποίηση αυτού του αναδυόμενου *ιστού των πραγμάτων* το «κλειδί» είναι οι μηχανές αναζήτησης, καθώς αυτές συνδέουν τους χρήστες και τις εφαρμογές με τους πόρους που απαιτούνται για τη λειτουργία τους. Στην εικόνα που ακολουθεί, Εικόνα 1, αποτυπώνεται γραφικά ο κομβικός ρόλος που παίζουν οι μηχανές αναζήτησης.

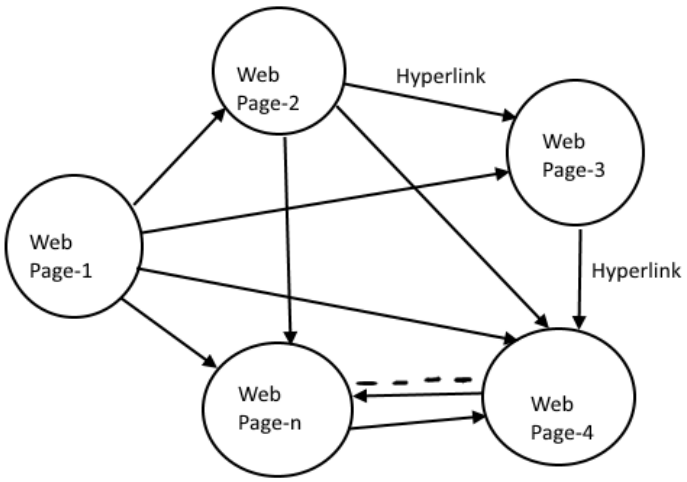


Εικόνα 1: Γραφική αναπαράσταση του ρόλου των μηχανών αναζήτησης στο *Web Of Things* [82].

Από τα παραπάνω, γίνεται αντιληπτό πως κάθε φάση εξέλιξης του διαδικτύου φέρνει την παγκόσμια ερευνητική κοινότητα αντιμέτωπη με νέες προκλήσεις. Είναι επόμενο πως αυτές οι νέες προκλήσεις αφορούν και το ζήτημα της οργάνωσης και διαχείρισης των σελίδων διαδικτύου, το οποίο μας απασχολεί στο πλαίσιο της παρούσας διατριβής. Οι προκλήσεις αυτές θίγονται αναλυτικότερα στην τελευταία Ενότητα αυτού του Κεφαλαίου (2.7. Προκλήσεις και Δυσκολίες στη Διαχείριση).

2.3. Δομή Παγκόσμιου Ιστού

Με τον ερχομό του *social web* και του *web of things*, εκτοξεύτηκε ο όγκος των δεδομένων που υπάρχουν στον παγκόσμιο ιστό. Μάλιστα, ο όγκος αυτός αυξάνεται συνεχώς και με τρόπο εκθετικό. Έτσι, προέκυψε η καθιέρωση του όρου *big data* [40]. Όσον αφορά τη δομή του παγκόσμιου ιστού, αυτή έχει εξελιχθεί από ένα συνολικό πληροφοριακό πεδίο σε ένα άλλο, όπου τόσο τα έγγραφα όσο και τα δεδομένα είναι διασυνδεδεμένα μεταξύ τους. Σε αυτή την εξέλιξη συνέβαλε ένα σύνολο βέλτιστων πρακτικών για την δημοσίευση και διασύνδεση δομημένων δεδομένων στον ιστό, γνωστά ως *Διασυνδεδεμένα Δεδομένα* (*Linked Data*). Στην Εικόνα 2 φαίνεται η γραφική αναπαράσταση της δομής του παγκόσμιου ιστού ως γράφου.



Εικόνα 2: Η δομή του παγκόσμιου ιστού ως γράφου [4].

Τα διασυνδεδεμένα δεδομένα συνίστανται στην χρήση του Ιστού για τη δημιουργία τυποποιημένων συνδέσμων ανάμεσα σε δεδομένα προερχόμενα από διαφορετικές

πηγές. Από μια πιο τεχνική σκοπιά, ο όρος «διασυνδεδεμένα δεδομένα» αναφέρεται σε δεδομένα δημοσιευμένα στον Ιστό με τέτοιο τρόπο, ώστε να είναι *μηχαναγνώσιμα* (machine-readable), η σημασία τους να ορίζεται με ρητό τρόπο, να συνδέονται με άλλα εξωτερικά σύνολα δεδομένων και να μπορούν με τη σειρά τους να χρησιμοποιηθούν από άλλα δεδομένα. Τα διασυνδεδεμένα δεδομένα συνίστανται ουσιαστικά σε μια συλλογή αλληλοσυσχετισμένων συνόλων δεδομένων, δημοσιευμένων στον Ιστό [17].

Παράλληλα, υπάρχουν ορισμένοι κανόνες που διέπουν τη δημοσίευση των διασυνδεδεμένων δεδομένων [11] και θα ήταν χρήσιμο να αναφερθούν εδώ, καθώς σχετίζονται με πλευρές της δομής του διαδικτύου και των σελίδων του. Πρόκειται για κανόνες που είναι γνωστοί ως «αρχές των διασυνδεδεμένων δεδομένων», και παρέχουν τη βασική μεθοδολογία για τη δημοσίευση και τη διασύνδεση δεδομένων, αξιοποιώντας τις υποδομές που παρέχει το διαδίκτυο, υιοθετώντας την αρχιτεκτονική του και ακολουθώντας τα πρότυπά του.

1. Μέσω των URIs προσδιορίζονται πράγματα/οντότητες/αντικείμενα.
2. Μέσω των HTTP URIs, οι χρήστες να μπορούν να αναφερθούν σε αυτά τα πράγματα/οντότητες/αντικείμενα και να μπορούν να τα αναζητήσουν.
3. Βλέποντας ένα URI, θα πρέπει να του παρέχονται χρήσιμες πληροφορίες για το περιεχόμενο, γι’ αυτό χρησιμοποιούμε τα πρότυπα (RDF, SPARQL).
4. Συμπεριλαμβάνονται σύνδεσμοι προς άλλα σχετικά URIs, ώστε οι χρήστες να μπορούν να ανακαλύψουν περισσότερες πληροφορίες.

Από πλευράς τεχνολογικής υποστήριξης, τα διασυνδεδεμένα δεδομένα βασίζονται σε δύο τεχνολογίες, ***U**niform **R**esource **I**dentifiers* (URIs) [14] και ***H**yper**T**ext **T**ransfer **P**rotocol* (HTTP) [13] οι οποίες είναι θεμελιώδεις για το διαδίκτυο. Καθώς τα ***U**niform **R**esource **L**ocators* (URLs) έχουν γίνει γνωστά σαν διευθύνσεις για τα τεκμήρια και άλλες οντότητες/αντικείμενα που βρίσκονται στο διαδίκτυο, τα URIs παρέχουν ένα πιο γενικό μέσο για την ταυτοποίηση κάθε οντότητας που υπάρχει στον κόσμο. Τα URIs και τα HTTPs συμπληρώνονται από την τεχνολογία των RDFs. Ενώ η HTML παρέχει ένα μέσο δόμησης και διασύνδεσης των τεκμηρίων στο διαδίκτυο, η RDF παρέχει ένα γενικό, βασισμένο στο γράφο μοντέλο δεδομένων, σύμφωνα με το οποίο

δομούνται και διασυνδέονται τα δεδομένα που περιγράφουν πράγματα/αντικείμενα/οντότητες του κόσμου [10].

Τα διασυνδεδεμένα δεδομένα αποτελούν ένα από τα είδη των δεδομένων που μας απασχολούν στην παρούσα διατριβή, καθώς αυτά καλούμαστε να διαχειριστούμε με σκοπό την κατηγοριοποίησή τους. Αυτά, τα διασυνδεδεμένα δεδομένα, μας ενδιαφέρουν από πλευράς περιεχομένου, δομής, οργάνωσης, συχνότητας επικαιροποίησής τους και άλλων. Το χαρακτηριστικό εκείνο που τα κάνει να διαφοροποιούνται από τα υπόλοιπα, είναι η διασυνδεσιμότητά τους, δηλαδή το γεγονός ότι η πληροφορία είναι διάσπαρτη και μάλιστα δεν μπορεί να εντοπιστεί χειρωνακτικά. Η ιδιαιτερότητά τους αυτή είναι και ο λόγος για τον οποίο δεν έχει ακόμα λυθεί το ζήτημα της οργάνωσης και γενικότερα της διαχείρισης των διασυνδεδεμένων δεδομένων. Ωστόσο, τα τελευταία χρόνια, πλήθος ερευνητών καταπιάνονται με το ζήτημα αρχικά της επεξεργασίας των διασυνδεδεμένων δεδομένων και κατά συνέπεια αυτό της διαχείρισης και οργάνωσής τους.

2.4. Δυναμικότητα

Παραδοσιακά, τα δεδομένα στον παγκόσμιο ιστό εμφανίζονται σχετικώς στατικά, αδόμητα, και με τον άνθρωπο να μπορεί να τα ελέγξει και να τα επεξεργαστεί. Καθώς το διαδίκτυο εξελίσσεται όμως, το περιεχόμενό του χαρακτηρίζεται δυναμικό με δύο τρόπους. Από τη μία εμπλουτίζεται και μετασχηματίζεται διαρκώς, και από την άλλη μεταβάλλεται με απρόβλεπτο τρόπο, σε απρόβλεπτο βαθμό και με απρόβλεπτο ρυθμό. Στο πλαίσιο της παρούσας διατριβής, μας απασχολούν οι αλλαγές που μπορεί να εμφανίζουν οι ίδιες σελίδες στο περιεχόμενό τους ή/και στη δομή τους.

Ιδιαίτερα τα τελευταία χρόνια, οι περιεχόμενες στον παγκόσμιο ιστό πληροφορίες μεταβάλλονται, λιγότερο ή περισσότερο. Αυτό συμβαίνει απρόβλεπτα, δηλαδή λιγότερο ή περισσότερο συχνά, και με ταχύτητα, ακόμα και μέσα σε λίγα λεπτά ή και δευτερόλεπτα [59]. Αυτό σημαίνει ότι ο δημιουργός τους μπορεί να ανανεώσει ή να διαγράψει το περιεχόμενό των σελίδων διαδικτύου ή μέρος αυτού, να το εμπλουτίσει ή και να αλλάξει δομικά χαρακτηριστικά της σελίδας, όπως είναι το URL της. Ο παράγοντας που κάνει ακόμα πιο σύνθετο το γεγονός αυτό, είναι πως η αλλαγές

μπορεί να πραγματοποιούνται ανομοιόμορφα. Αυτό σημαίνει ότι η κάθε σελίδα μπορεί να αλλάζει με το δικό της ρυθμό, ο οποίος δεν είναι σταθερός, και σε διαφορετικό βαθμό κάθε φορά ή σε διαφορετικό βαθμό σε σχέση με άλλες σελίδες. Σε αυτή την παρατήρηση είναι που συνοψίζεται και το ενδιαφέρον των ερευνητών που μελετούν την καλύτερη και αποτελεσματικότερη διαχείριση του περιεχομένου του διαδικτύου, αφού αυτά τα χαρακτηριστικά δημιουργούν επιπλέον δυσκολίες κατά την υιοθέτηση «παραδοσιακών» τεχνικών ανάλυσης κειμένου για τη διαχείριση των σελίδων διαδικτύου [23] [40].

Διατρέχοντας τη βιβλιογραφία, αντιλαμβανόμαστε πως ο εντοπισμός του μεγέθους των αλλαγών των σελίδων διαδικτύου, το μοτίβο των αλλαγών, η συχνότητα εμφάνισής τους και η φύση τους κεντρίζουν το ενδιαφέρον ερευνητών πολλών πεδίων. Παράλληλα, υπάρχουν *συστήματα ανίχνευσης των αλλαγών και ειδοποίησης* (Change Detection and Notification Systems – CDN systems), όπως το Visualping [V] και το Wachete [X], τα οποία παρέχουν διάφορες επιλογές παρακολούθησης. Για παράδειγμα, μπορεί να επιλέξει κάποιος αν θέλει να παρακολουθεί μέρος μιας σελίδας ή ολόκληρη, ορισμένες σελίδες ενός ιστοτόπου ή ολόκληρο τον ιστότοπο. Παράλληλα, μπορεί να οριστεί η συχνότητα του ελέγχου (ωριαία/ημερήσια/εβδομαδιαία/μηναία βάση) και των ειδοποιήσεων (δύο φορές τη μέρα, μία φορά τη μέρα, μία φορά την εβδομάδα ή όταν εντοπίζεται κάποια αλλαγή) [47].

Είναι φανερό πως η ανίχνευση των αλλαγών που παρουσιάζει το περιεχόμενο του διαδικτύου, αποτελεί πλέον μέρος της έρευνας για την αποτελεσματική διαχείριση του διαδικτυακού περιεχομένου. Έτσι, απασχολεί και εμάς στο πλαίσιο της παρούσας διατριβής, όπου μελετάμε την κατηγοριοποίηση των διαδικτυακών σελίδων, καθώς ζητούμενο είναι η κατηγοριοποίηση να παραμένει πάντα επικαιροποιημένη.

2.5. Ποικιλομορφία και Ετερογένεια

Τα δεδομένα διαδικτύου χαρακτηρίζονται, επίσης, από ετερογένεια και ποικιλομορφία. Αυτό προκύπτει από την αναπαράσταση των δεδομένων που δημιουργούνται από διαφορετικές πηγές, και άρα έχουν διαφορετικά

χαρακτηριστικά. Για παράδειγμα, οι σελίδες που σχετίζονται με το ηλεκτρονικό εμπόριο περιλαμβάνουν δομημένα δεδομένα, τα αρχεία καταγραφής διακομιστών ιστού (web server log) ημι-δομημένα δεδομένα, ενώ οι σελίδες κοινωνικής δικτύωσης περιλαμβάνουν δεδομένα που είναι αδόμητα. Μάλιστα, όσο ο παγκόσμιος ιστός αναπτύσσεται, δημιουργούνται σελίδες που εμφανίζουν χαρακτηριστικά και των τριών αυτών κατηγοριών [20]. Αυτό έχει σαν αποτέλεσμα τη δημιουργία ολοένα και μεγαλύτερων απαιτήσεων για τη διαλειτουργικότητα των διαφορετικών εφαρμογών και υπηρεσιών που παρέχονται μέσω διαδικτύου. Παράλληλα, υπάρχουν πολλές και διαφορετικές γλώσσες προγραμματισμού μέσω των οποίων δημιουργούνται τα διαδικτυακά δεδομένα, η καθεμιά εκ των οποίων έχει τους δικούς της κανόνες σύνταξης και το δικό της λεξιλόγιο. Ορισμένες εξ αυτών είναι οι xml, html, css, java. Εκτός από την ποικιλομορφία κατά την δημιουργία των δεδομένων διαδικτύου, παρατηρείται ολοένα και μεγαλύτερη ποικιλομορφία στο περιεχόμενο των σελίδων, το οποίο μπορεί να έχει τη μορφή εικόνας, κειμένου, ήχου, βίντεο ή συνδυασμό ορισμένων ή και όλων. Τα τελευταία χρόνια, υπάρχει ακόμα μια μορφή δεδομένων που συναντάται συχνά στο διαδίκτυο, τα *δεδομένα κοινωνικής δικτύωσης* (social data).

Προκειμένου να αντιμετωπιστεί η ετερογένεια στην τυπολογία των σελίδων διαδικτύου, ερευνητές παρατηρούν πως σελίδες που ανήκουν σε κάποια συγκεκριμένη κατηγορία, έχουν ορισμένες ομοιότητες στη δομή τους. Αυτή τη γενική δομή των σελίδων μπορεί να τη συμπεράνει κάποιος από τη θέση των συνδέσμων, του κειμένου και των εικόνων. Παράλληλα, υπάρχουν μερικά χαρακτηριστικά, όπως είναι η αναλογία/ο λόγος της ποσότητας των συνδέσμων προς αυτή του κειμένου, τα οποία εμφανίζονται σε όλες τις κατηγορίες των εγγράφων, αλλά σε διαφορετικό βαθμό, και υπάρχουν και κάποια άλλα χαρακτηριστικά που συναντώνται μόνο σε ορισμένες κατηγορίες ιστοσελίδων. Βασισμένοι σε αυτή τη θεώρηση, οι [7] κατηγοριοποιούν τις ιστοσελίδες στις εξής βασικές κατηγορίες:

- *Πληροφοριακές σελίδες (informational pages)*
- *Σελίδες αναζήτησης (research pages)*
- *Προσωπικές αρχικές σελίδες (personal home pages)*

Στην προσπάθειά να καλυφθούν τα κενά των προηγούμενων μελετών, στην έρευνα των [39] κατηγοριοποιούνται σαφώς και πλήρως τα πληροφοριακά αιτήματα των χρηστών, λαμβάνοντας υπόψη τόσο το ίδιο το αίτημα, μιας και αυτό αποτελεί το δεδομένο-κλειδί του σκοπού μιας αναζήτησης, όσο και άλλους παράγοντες, όπως είναι η σελίδα την οποία επιλέγει τελικώς ο χρήστης να επισκεφτεί θεωρώντας την καταλληλότερη και σχετικότερη με το αίτημά του. Αποτέλεσμα αυτής της μελέτης είναι η κατάταξη των πληροφοριακών αιτημάτων στις παρακάτω κατηγορίες με τα αντίστοιχα χαρακτηριστικά:

- *Αναζήτηση με σκοπό την πληροφόρηση (informational searching)*: πρόκειται για αναζήτηση με σκοπό τον εντοπισμό περιεχομένου που σχετίζεται με ένα ορισμένο θέμα, προκειμένου ο χρήστης να ικανοποιήσει την πληροφοριακή του ανάγκη. Το περιεχόμενο μπορεί να είναι πολλών μορφών, συμπεριλαμβανομένων κειμένου, δεδομένων, τεκμηρίων και οπτικοακουστικού υλικού. Η ανάγκη του χρήστη μπορεί να είναι από κάτι πολύ συγκεκριμένο, μέχρι και κάτι πολύ γενικό.
- *Αναζήτηση με σκοπό την πλοήγηση (navigational searching)*: πρόκειται για αναζήτηση με σκοπό τον εντοπισμό μιας ορισμένης ιστοσελίδας, μέσα στην οποία ο χρήστης επιθυμεί να πλοηγηθεί. Η ιστοσελίδα αυτή μπορεί να είναι είτε ενός ατόμου είτε ενός οργανισμού. Σε αυτή την περίπτωση, κατά κανόνα, ο χρήστης μπορεί ήδη να γνωρίζει την ύπαρξη της ιστοσελίδας που αναζητά ή μπορεί απλώς να θεωρεί ότι υπάρχει.
- *Αναζήτηση με σκοπό τη συναλλαγή (transactional searching)*: πρόκειται για αναζήτηση με σκοπό τον εντοπισμό μιας ιστοσελίδας μέσω της οποίας ο χρήστης θα μπορέσει να αποκτήσει ένα προϊόν, να πραγματοποιήσει κάποια συναλλαγή ή να εξυπηρετηθεί στο πλαίσιο μιας υπηρεσίας.

Σε συνέχεια της κατηγοριοποίησης που αναφέρεται πιο πάνω, οι [38] προσδιορίζουν και τα αντίστοιχα χαρακτηριστικά των αιτημάτων που υποβάλλουν οι χρήστες στο πλαίσιο των αναζητήσεών τους με βάση τις παραπάνω κατηγορίες:

- *Πληροφοριακά αιτήματα (informational queries)*:
    - Τα ερωτήματα μπορεί να περιλαμβάνουν ερωτηματικές λέξεις/φράσεις.

- Τα ερωτήματα μπορεί να περιλαμβάνουν όρους φυσικής γλώσσας.
- Τα ερωτήματα μπορεί να περιλαμβάνουν όρους πληροφόρησης, π.χ. κατάλογος κ.λπ.
- Τα ερωτήματα μπορεί να είναι συνέχεια προηγούμενων ερωτημάτων.
- Τα ερωτήματα μπορεί να υποβάλλονται αφότου ο χρήστης έχει δει ορισμένες σελίδες που του δίνονται σαν αποτέλεσμα της αναζήτησής του.
- Τα ερωτήματα περιλαμβάνουν τουλάχιστον δύο λέξεις.
- Τα ερωτήματα αυτά δεν έχουν κάποιο από τα χαρακτηριστικά των άλλων δύο κατηγοριών ερωτημάτων.

- *Αιτήματα πλοήγησης (navigational queries)*:
  - Τα ερωτήματα μπορεί να περιέχουν το όνομα μιας εταιρίας/επιχείρησης, ενός οργανισμού ή ενός ατόμου .
  - Τα ερωτήματα μπορεί να περιλαμβάνουν μέρος διαδικτυακής διεύθυνσης (domain suffix).
  - Τα ερωτήματα μπορεί να έχουν το «διαδίκτυο» σαν πηγή.
  - Το μήκος των ερωτημάτων συνήθως δεν ξεπερνά τις 3 λέξεις.
  - Ο χρήστης συνήθως επιλέγει να προβάλει το πρώτο από τα αποτελέσματα που του επιστρέφει η μηχανή αναζήτησης.
- *Αιτήματα συναλλαγής (transactional queries)*:
  - Τα ερωτήματα μπορεί να περιλαμβάνουν όρους που σχετίζονται με ταινίες, τραγούδια, στίχους, συνταγές, εικόνες, χιούμορ, πορνογραφία.
  - Τα ερωτήματα μπορεί να περιλαμβάνουν όρους για την «απόκτηση» ταινιών, τραγουδιών κ.λπ.
  - Τα ερωτήματα μπορεί να περιλαμβάνουν όρους για τη «λήψη» ταινιών, στίχων, λογισμικού κ.λπ.
  - Τα ερωτήματα μπορεί να σχετίζονται με μια εικόνα, ένα οπτικοακουστικό υλικό, ένα ηχητικό ντοκουμέντο.
  - Τα ερωτήματα μπορεί να περιλαμβάνουν όρους ψυχαγωγίας, π.χ. παιχνίδια.

- Τα ερωτήματα μπορεί να περιλαμβάνουν όρους «διάδρασης», π.χ. αγοράζω, συζητώ.
- Τα ερωτήματα μπορεί να περιλαμβάνουν κάποια επέκταση συμπιεσμένων αρχείων μουσικής, ταινιών, εικόνων κ.λπ., όπως είναι οι επεκτάσεις jpeg, zip κ.λπ.

Τα ευρήματα των παραπάνω μελετών αποτελούν τη βάση για τη δομική κατηγοριοποίηση στο πλαίσιο της προτεινόμενης μεθοδολογίας. Εκτενέστερη αναφορά γίνεται στη σχετική Ενότητα, 4.3.1. Πολυδιάστατη Κατηγοριοποίηση Σελίδων Διαδικτύου.

2.6. Αμφίβολη Ποιότητα Δεδομένων

Ο όρος «big data» καθιερώθηκε σχετικώς πρόσφατα, γεγονός που εξηγεί την έλλειψη ενιαίων και σαφώς ορισμένων κριτηρίων ποιότητας αυτών των δεδομένων. Ωστόσο, αυτό που ως βεβαιότητα αναφέρεται στη σχετική βιβλιογραφία είναι ότι η ποιότητα των δεδομένων δεν εξαρτάται μόνο από τα χαρακτηριστικά των ίδιων, αλλά και από το περιβάλλον δημιουργίας και το περιβάλλον χρήσης τους. Συνήθως τα πρότυπα ποιότητας, όπου υπάρχουν, αναπτύσσονται από την οπτική των δημιουργών των δεδομένων. Στο παρελθόν, οι χρήστες των δεδομένων ήταν και αυτοί που, αμέσως ή εμμέσως, τα δημιουργούσαν κιόλας. Αυτό διασφάλιζε και την ποιότητά τους. Όμως, στην εποχή των μεγάλου όγκου δεδομένων (big data), όπου οι πηγές αυτών των δεδομένων είναι πολλές και ποικίλες, και οι χρήστες δεν είναι απαραίτητα καί δημιουργοί τους, ο έλεγχος και η διασφάλιση ποιότητας των δεδομένων είναι ακόμα πιο δύσκολο να είναι αυστηρά [19].

Αυτό έχει σαν αποτέλεσμα να συναντούν οι χρήστες συχνά στον παγκόσμιο ιστό παραπλανητικές ή εξ ολοκλήρου ψευδείς πληροφορίες, να λαμβάνουν ανεπιθύμητη αλληλογραφία, να διαβάζουν ανορθόγραφα ή/και ασύντακτα κείμενα, ακόμα και να οδηγούνται σε ιστοτόπους που στόχο έχουν να αλλοιώσουν το λογισμικό των υπολογιστών τους. Αυτό ενισχύεται από τη δυνατότητα της ανωνυμίας που παρέχει το διαδίκτυο. Προέκταση αυτής της συνθήκης θα μπορούσαμε να πούμε πως είναι και το ηλεκτρονικό έγκλημα, αλλά και η συμβατικών μορφών εγκληματικότητα που

συχνά βρίσκει πρόσφορο έδαφος στο διαδίκτυο. Μάλιστα, ήδη από τη δεκαετία του 1990, εκτός από το διαδίκτυο που όλοι γνωρίζουμε και χρησιμοποιούμε, το λεγόμενο «Surface Web» του οποίου το περιεχόμενο -αξιόπιστο ή μη- είναι δημόσιο, υπάρχει και το «Deep Web», του οποίου το περιεχόμενο είναι ιδιωτικό, αλλά και το «Darknet», το οποίο είναι προσβάσιμο με τη χρήση συγκεκριμένου λογισμικού (TOR – The Onion Router), του οποίου οι χρήστες είναι αδύνατο να εντοπιστούν, και συνδέεται στενά με την παροχή πληροφοριών σχετικών της αυτοχειρίας [51].

Δεδομένου ότι σκοπός του παρόντος Κεφαλαίου είναι η αναφορά στα βασικά χαρακτηριστικά και τις ιδιαιτερότητες του παγκόσμιου ιστού, μένουμε περισσότερο στο θέμα των αμφίβολης ποιότητας των δεδομένων που αφορά η Ενότητα αυτή. Βασικό μέρος αυτών των δεδομένων αποτελούν οι ψευδείς ειδήσεις (fake news). Σχετικά με αυτά, ενδιαφέρον παρουσιάζει η «τρισδιάστατη» ανάγνωση του «οικοσυστήματος» των ψευδών ειδήσεων από ερευνητές [76], σύμφωνα με την οποία οι ψευδείς ειδήσεις εξαρτώνται από τρεις παραμέτρους: το περιεχόμενο, την κοινωνική διάσταση και την χρονική διάσταση. Η παράμετρος του περιεχομένου έχει να κάνει με τη συσχέτιση των ειδήσεων/δημοσιεύσεων με σχόλια ή/και σχετικές αναρτήσεις χρηστών σε κοινωνικά μέσα. Η κοινωνική διάσταση αφορά στις σχέσεις μεταξύ των δημιουργών/εκδοτών, διανομέων και χρηστών. Ενώ η χρονική διάσταση αντικατοπτρίζει την εξέλιξη των δημοσιεύσεων και των αναρτήσεων των χρηστών με την πάροδο του χρόνου. Μάλιστα, οι ίδιοι υποστηρίζουν πως η μελέτη και η εξέταση αυτών των παραμέτρων μπορεί να συμβάλει καθοριστικά στον εντοπισμό των ψευδών ειδήσεων.

Σε κάθε περίπτωση, ο εντοπισμός, ο περιορισμός και η εξάλειψη των αμφίβολης ποιότητας δεδομένων στο πλαίσιο του παγκόσμιου ιστού, είναι κάτι που κεντρίζει το ενδιαφέρον πολλών ερευνητών και επιστημόνων προερχόμενων από πολλούς και διαφορετικούς τομείς [90] [91].

2.7. Προκλήσεις και Δυσκολίες στη Διαχείριση

Από τα παραπάνω, φαίνεται πως η ίδια η φύση του παγκόσμιου ιστού είναι εκείνη που δημιουργεί και τις προκλήσεις για τη διαχείρισή του. Με άλλα λόγια, οι

ιδιαιτερότητες που χαρακτηρίζουν τα δεδομένα διαδικτύου και αναφέρονται πιο πάνω, είναι αυτές που προκαλούν δυσκολίες και ταυτόχρονα συνιστούν τις τρέχουσες προκλήσεις για τη διαχείρισή τους. Αυτό είναι και το κίνητρο ως αφετηρία της παρούσας διατριβής. Η σημασία ή ακόμα και η ανάγκη, θα μπορούσαμε να πούμε, της ερευνητικής μελέτης σχετικών ζητημάτων, έχοντας στόχο την καλύτερη διαχείριση και οργάνωση των διαδικτυακών δεδομένων, ενισχύεται από το γεγονός ότι το διαδίκτυο από βοηθητικό εργαλείο έχει μετατραπεί σε πρωταρχικό για τη ζωή του ανθρώπου. Αυτό αφορά όλες τις πλευρές της ανθρώπινης ζωής· απλές και σύνθετες.

Πιο συγκεκριμένα, και επιστρέφοντας στα πιο τεχνικά ζητήματα, όπως είναι η διαδικασία της αναζήτησης που είναι θεμελιώδης για το Web-of-Things (βλ. 2.2. Ιστορική Αναδρομή), γίνεται ολοένα και πιο απαιτητική λόγω της κινητικότητας των αντικειμένων, της προσωρινής παρουσίας και της συνεχούς ροής δεδομένων με μεταβαλλόμενες χωρικές και χρονικές ιδιότητες. Κατά συνέπεια, είναι απαραίτητη η αποτελεσματική ευρετηρίαση τόσο των ιστορικών δεδομένων όσο και των δεδομένων πραγματικού χρόνου. Η ερευνητική κοινότητα έχει αναπτύξει πολυάριθμες τεχνικές και μεθόδους για την αντιμετώπιση αυτών των προκλήσεων, οι οποίες αφορούν τις βασικές αρχές του παγκόσμιου ιστού, την αναπαράσταση των δεδομένων, αλλά και το περιεχόμενο υπό αναζήτηση. Σχετικά με τις βασικές αρχές που διέπουν τον παγκόσμιο ιστό, οι παραδοσιακές τεχνικές που καθιστούν εφικτή την αναζήτηση και ανάκτηση δεδομένων, όπως είναι η ευρετηρίαση, η ομαδοποίηση και η ταξινόμηση, δεν μπορούν να εφαρμοστούν κατ’ ευθείαν. Χρειάζεται να προσαρμοστούν για να είναι ικανές να υποστηρίξουν την αναζήτηση και στο Web-of-Things.

Την ίδια στιγμή, παραδοσιακά, ο σημασιολογικός ιστός στοχεύει στην παροχή σημασιολογίας και διαλειτουργικότητας για κάθε τύπο πόρων στον Ιστό, για την οικοδόμηση ενός δικτύου δεδομένων. Οι κοινότητες που ασχολούνται με το θέμα της αναζήτησης έχουν αναγνωρίσει τη σημασία των τεχνικών του σημασιολογικού ιστού (π.χ. τεχνικές αναπαράστασης της γνώσης και αυτοματοποιημένη συλλογιστική για την εξαγωγή και δημιουργία νέας γνώσης), στην πραγματοποίηση έξυπνων

υπηρεσιών με συνδεδεμένα αντικείμενα στο IoT (Internet of Things) και στο WoT (Web of Things). Χρειάζεται όμως να διερευνηθούν οι τεχνικές αναζήτησης από την πλευρά της αναπαράστασης δεδομένων γνώσης, με ιδιαίτερη έμφαση στη χρήση των τεχνολογιών του Σημασιολογικού Ιστού, δηλαδή των Συνδεδεμένων Δεδομένων και των σημασιολογικών δεδομένων ροής. Κάτι αντίστοιχο, δηλαδή η απαραίτητη προσαρμογή, χρειάζεται να γίνει και για τις τεχνικές αναζήτησης αναφορικά με το περιεχόμενο, δεδομένου ότι η ετερογένεια στο περιεχόμενο συνεπάγεται και ετερογένεια στις μεθόδους αναζήτησης. Πρακτικά, η ανάγκη αυτή συνδέεται άμεσα και άρρηκτα με την ανάγκη για βελτιωμένες τεχνικές διαχείρισης και οργάνωσης του διαδικτύου.

Στο πλαίσιο αυτό, οι ερευνητές καλούνται να αντιμετωπίσουν δυσκολίες, που παραδοσιακά κεντρίζουν το ενδιαφέρον τους κ αντιμετωπίζονται επιτυχώς, όμως την ίδια στιγμή παραμένουν ως προκλήσεις, καθώς με την εξέλιξη του παγκόσμιου ιστού, μεταλλάσσονται και εξελίσσονται μαζί του. Αυτές συνοψίζονται στα εξής [19].

- Η ***ποικιλομορφία των πηγών δεδομένων*** συνεπάγεται και ποικιλομορφία στον τύπο των δεδομένων, πιο σύνθετη δομή, με αποτέλεσμα να γίνεται πιο απαιτητική και η διαλειτουργικότητα μεταξύ τους. Αυτό, κάνει πιο δύσκολη την ανάκτησή τους από τις μηχανές αναζήτησης ως απάντηση σε ένα αίτημα αναζήτησης χρήστη, καθώς αυτή η ποικιλομορφία κάνει πιο σύνθετη την αυτοματοματοποιημένη διαχείρισή τους.
- Ο ***όγκος των διαθέσιμων δεδομένων*** είναι τεράστιος και είναι δύσκολο να εκτιμηθεί η ποιότητά τους εντός ενός εύλογου χρονικού διαστήματος. Κι αυτό γιατί απαιτείται πρώτα η συλλογή τους, ύστερα πρέπει να «καθαριστούν», να προσαρμοστούν ώστε να διαλειτουργούν και τελικώς να προκύψουν τα υψηλής ποιότητας δεδομένα που είναι απαραίτητα. Η δυσκολία αυξάνεται αν συνυπολογιστεί ο χρόνος που χρειάζονται τα αδόμητα δεδομένα να εναρμονιστούν με δομημένους τύπους, ώστε να είναι δυνατή η περαιτέρω επεξεργασία των δεδομένων. Έτσι, παρότι η αυτοματοποιημένη επεξεργασία των διαδικτυακών δεδομένων απασχολεί την ερευνητική κοινότητα από τότε που εμφανίστηκε το διαδίκτυο, οι απαιτήσεις που δημιουργούνται στον τομέα αυτό, γίνονται ολοένα και πιο προκλητικές.

- Τα ***δεδομένα αλλάζουν*** πολύ γρήγορα και με τρόπο απρόβλεπτο και «άναρχο», με αποτέλεσμα να «ξεπερνώνται» σύντομα, χάνοντας μέρος της πληροφοριακής και όχι μόνο αξίας τους. Το γεγονός αυτό ανεβάζει τον πήχη των απαιτήσεων για την επεξεργασία τους, καθώς, για να έχουμε πάντα επικαιροποιημένα αποτελέσματα κατά την ανάκτηση πληροφοριών και όχι μόνο, τα απαιτούμενα δεδομένα θα πρέπει να συλλεχθούν ή/και να υποστούν επεξεργασία εγκαίρως - εννοώντας σε πραγματικό χρόνο. Παράλληλα, όπως φαίνεται και από την ιστορική αναδρομή που γίνεται σε προηγούμενη ενότητα (2.2. Ιστορική Αναδρομή) δεν μεταβάλλεται μόνο το περιεχόμενο του διαδικτύου, αλλά και το διαδίκτυο ως οντότητα, καθώς εξελικτικά αλλάζει η χρήση του.
- Υπάρχει ***έλλειψη ενιαίων/καθολικών προτύπων διασφάλισης ποιότητας*** των δεδομένων. Το γεγονός αυτό, ιδίως αν συνδυαστεί με την ποικιλομορφία των πηγών δεδομένων και τον όγκο τους, ενισχύει ακόμα περισσότερο την ανάγκη για ολοένα και πιο γρήγορες και αποτελεσματικές τεχνικές επεξεργασίας των διαδικτυακών δεδομένων.

Τα παραπάνω αποτελούν την αφετηρία, τη βάση και το κίνητρο για την εκπόνηση της παρούσας διατριβής. Έχοντας στόχο την αυτοματοποιημένη διαχείριση και οργάνωση των δυναμικών δεδομένων, σχεδιάζεται μια τεχνική αυτοματοποιημένης κατηγοριοποίησης σελίδων διαδικτύου, η οποία είναι συνδυαστική/σύνθετη, στηρίζεται σε ευρήματα εξέχοντων πρότερων μελετών, και αξιοποιεί απλά και κοινά στοιχεία με τρόπο νέο.

ΚΕΦΑΛΑΙΟ 3: Τεχνικές Κατηγοριοποίησης Διαδικτυακών Δεδομένων

3.1. Εισαγωγή

Η *κατηγοριοποίηση δεδομένων* (data classification) είναι μια από τις τεχνικές *εξόρυξης γνώσης* (data mining), που εξυπηρετεί την εξαγωγή πληροφοριών και προτύπων χρήσιμων στη λήψη αποφάσεων. Παράδειγμα κατηγοριοποίησης από την καθημερινή ζωή, αποτελεί ο εντοπισμός, και στη συνέχεια διαχωρισμός της *εισερχόμενης ανεπιθύμητης ηλεκτρονικής αλληλογραφίας* (spam e-mails), ο καθορισμός των ατόμων που υγειονομικά χαρακτηρίζονται ως μέλη ευπαθών ομάδων κ.α. Εύκολα γίνεται αντιληπτό το γεγονός ότι όσο αυξάνεται ο όγκος των δεδομένων προς επεξεργασία, τόσο αυξάνονται οι απαιτήσεις κατά τη διαδικασία της επεξεργασίας, με αποτέλεσμα την αυτοματοποίηση των τεχνικών αυτών ολοένα και περισσότερο. Με ανάλογο τρόπο εξελίσσεται και αναπτύσσεται η απαίτηση για πιο αποτελεσματικές τεχνικές κατηγοριοποίησης διαδικτυακών δεδομένων, αφού τα τελευταία αυξάνονται ιλιγγιωδώς και χαρακτηρίζονται από ορισμένες ιδιαιτερότητες που ανεβάζουν επιπλέον τον πήχη για αποτελεσματικότερη διαχείριση και οργάνωσή τους. Έτσι, η αυτόματη κατηγοριοποίηση διαδικτυακών και όχι μόνο δεδομένων εξακολουθεί να κεντρίζει το ενδιαφέρον της διεθνούς ερευνητικής κοινότητας.

Στο πλαίσιο της παρούσας διατριβής, τα διαδικτυακά δεδομένα που μας απασχολούν είναι οι σελίδες διαδικτύου. Σύμφωνα με τη βιβλιογραφία [63], το πρόβλημα της κατηγοριοποίησής τους μπορεί, σε ένα πρώτο επίπεδο, να διαιρεθεί στα εξής επιμέρους προβλήματα:

(i) *θεματική κατηγοριοποίηση* (subject classification), η οποία αφορά το θέμα που πραγματεύεται η σελίδα (π.χ. τέχνη, αθλητισμός, επιστήμη).

(ii) *λειτουργική κατηγοριοποίηση* (functional classification), η οποία αφορά το ρόλο που παίζει η σελίδα (προσωπική ιστοσελίδα, σελίδα μαθήματος),

(iii) *κατηγοριοποίηση συναισθήματος* (sentiment classification), η οποία αφορά τη γνώμη που παρουσιάζεται σε μια ιστοσελίδα (π.χ. γνώμη που έχει ο γράφων γύρω από ένα συγκεκριμένο ζήτημα).

Στο πλαίσιο της παρούσας διατριβής, στο μεγαλύτερο μέρος της σχετικής έρευνας που μελετάμε συναντάμε τεχνικές αυτόματης κατηγοριοποίησης διαδικτυακών σελίδων, που κυρίως στηρίζονται είτε στο περιεχόμενο (θεματική κατηγοριοποίηση) είτε στη δομή τους (λειτουργική κατηγοριοποίηση). Έτσι, στις Ενότητες που ακολουθούν, γίνεται επισκόπηση των τεχνικών αυτόματης κατηγοριοποίησης διαδικτυακών σελίδων, που αποτέλεσαν τη βάση και αφετηρία της παρούσας δουλειάς.

3.2. Τεχνικές Κατηγοριοποίησης Σελίδων Διαδικτύου

Ορισμός: Κατηγοριοποίηση είναι η διαδικασία κατά την οποία μια τεχνική καθίσταται ικανή να προβλέψει την κατηγορία (κλάση) ενός στοιχείου επιλέγοντας μεταξύ καθορισμένων τιμών. Πρόκειται για μια διαδικασία που ολοκληρώνεται μέσα από δύο βήματα. Κατά τη διάρκεια του πρώτου βήματος, εφαρμόζοντας τον αλγόριθμο κατηγοριοποίησης στο σύνολο δεδομένων εκπαίδευσης, δημιουργείται το μοντέλο κατηγοριοποίησης. Στη συνέχεια, κατά το δεύτερο βήμα, το εξαχθέν μοντέλο δοκιμάζεται σε ένα προκαθορισμένο σύνολο δοκιμαστικών δεδομένων, και μετρώνται η απόδοση και η ακρίβεια του μοντέλου [56].

Η έρευνα γύρω από το ζήτημα αυτό έχει εστιαστεί προς τρεις βασικές κατευθύνσεις. Αυτές είναι η *στατιστική* (statistics), η *μηχανική μάθηση* (machine learning) και τα *νευρωνικά δίκτυα* (neural networks). Οι προσεγγίσεις που βασίζονται στη στατιστική χαρακτηρίζονται από την ύπαρξη ενός βασικού μοντέλου πιθανοτήτων που δουλεύει με ακρίβεια, και το οποίο παρέχει την πιθανότητα να βρίσκεται σε κάθε κλάση το εκάστοτε στοιχείο προς κατηγοριοποίηση. Από την άλλη, τα μοντέλα μηχανικής μάθησης βασίζονται σε λογικές ή δυαδικές λειτουργίες, και μαθαίνουν μια εργασία μέσα από μία σειρά παραδειγμάτων. Με άλλα λόγια, στοχεύουν στη δημιουργία εκφράσεων κατηγοριοποίησης αρκετά απλών, που να «μιμούνται» τον ανθρώπινο συλλογισμό επαρκώς, ώστε να παρέχουν τη διαδικασία λήψης αποφάσεων. Τέλος,

το πεδίο των νευρωνικών δικτύων έχει προκύψει από διαφορετικές αφετηρίες, των οποίων το εύρος κυμαίνεται από την κατανόηση και μίμηση του ανθρώπινου εγκεφάλου έως ευρύτερα θέματα αντιγραφής των ανθρώπινων ικανοτήτων, όπως είναι η ομιλία. Γενικά, τα νευρωνικά δίκτυα αποτελούνται από στρώματα διασυνδεδεμένων κόμβων, όπου κάθε κόμβος παράγει μια μη-γραμμική συνάρτηση των δεδομένων εισόδου, τα οποία μπορεί να προέρχονται από άλλους κόμβους ή απευθείας από τα αρχικά δεδομένα εισόδου, ενώ ορισμένοι κόμβοι ταυτίζονται με την έξοδο του δικτύου.

Μιλώντας πιο συγκεκριμένα για την κατηγοριοποίηση των σελίδων διαδικτύου, που είναι και το αντικείμενο μελέτης της παρούσας διατριβής, η κατηγοριοποίηση είναι η διαδικασία της απόδοσης μιας σελίδας σε μία ή περισσότερες προκαθορισμένες κατηγορίες. Παραδοσιακά, η κατηγοριοποίηση των σελίδων διαδικτύου αντιμετωπίζεται ως ένα πρόβλημα εποπτευόμενης μάθησης [74], όπου ένα σύνολο κατηγοριοποιημένων δεδομένων χρησιμοποιούνται για την «εκπαίδευση» ενός κατηγοριοποιητή, ο οποίος μπορεί να εφαρμοστεί για την κατηγοριοποίηση άλλων δεδομένων.

Στη συνέχεια, παρουσιάζονται ορισμένοι από τους πιο διαδεδομένους αλγόριθμους αυτόματης κατηγοριοποίησης κειμένων, στους οποίους βασίζονται με τη σειρά τους οι παραδοσιακές τεχνικές κατηγοριοποίησης σελίδων διαδικτύου. Ωστόσο, κρίνεται σκόπιμο να διαχωριστούν νωρίτερα τα δύο πεδία, καθώς η κατηγοριοποίηση σελίδων διαδικτύου διαφέρει από αυτή των κειμένων, λόγω της φύσης των πρώτων, όπως αναφέρεται και σε προηγούμενο Κεφάλαιο (**ΚΕΦΑΛΑΙΟ 2:** Χαρακτηριστικά και Ιδιαιτερότητες Διαδικτύου). Αρχικώς, οι παραδοσιακές μέθοδοι κατηγοριοποίησης κειμένων βρίσκουν εφαρμογή κυρίως σε δομημένα έγγραφα γραμμένα με συνοχή και δομική συνέπεια, ενώ τα περιεχόμενο του διαδικτύου δεν έχει αυτό το χαρακτήρα. Ύστερα, οι σελίδες διαδικτύου αποτελούν ημι-δομημένα έγγραφα γραμμένα σε HTML γλώσσα, όπου ακόμα κι αν υπάρχουν επισημειώσεις κατά τη δημιουργία τους, αυτές είναι πολύ περιορισμένες για να ανταποκριθούν στις ανάγκες της κατηγοριοποίησης. Τέλος, οι διαδικτυακές σελίδες υπάρχουν εντός ενός υπερ-κειμένου, με συνδέσεις από και προς αυτές. Ιδίως το τελευταίο, η διασυνδεσιμότητα

των σελίδων διαδικτύου είναι βασικό χαρακτηριστικό του Ιστού που απουσιάζει από το πρόβλημα κατηγοριοποίησης παραδοσιακών κειμένων. Επομένως, το θέμα της κατηγοριοποίησης σελίδων διαδικτύου είναι, εκτός από σημαντικό, και διαφορετικό στο βάθος του από αυτό της κατηγοριοποίησης κειμένων [63].

Από μία άλλη οπτική, οι ίδιοι οι περιορισμοί των μεθόδων κατηγοριοποίησης κειμένων όταν οι τελευταίοι εφαρμόζονται για την κατηγοριοποίηση των σελίδων διαδικτύου, μας οδηγούν και στη «διέξοδο» από αυτούς. Συγκεκριμένα, βλέποντας πως οι παραδοσιακές μέθοδοι κατηγοριοποίησης κειμένου είναι πιο αποτελεσματικές σε απολύτως δομημένα δεδομένα, οι ερευνητές οδηγούνται στην αξιοποίηση και ανάπτυξη μεθόδων ημι-εποπτευόμενης μάθησης [36].

3.2.1. *k*-Nearest Neighbors (*k*-NN)

Η προσέγγιση του *k*-Nearest Neighbor (*k*-πλησιέστερος γείτονας) είναι σχετικώς απλή: δεδομένου ενός *δοκιμαστικού εγγράφου d* (test document *d*), το σύστημα εντοπίζει *τον k-πλησιέστερο γείτονα* μεταξύ των *εγγράφων εκπαίδευσης* (training documents), και αξιοποιεί τις *κλάσεις* τού *k*-πλησιέστερου γείτονα για να ελέγξει τη *βαρύτητα* των υποψήφιων κλάσεων. Η βαρύτητα των κλάσεων του κάθε πλησιέστερου γείτονα-εγγράφου υπολογίζεται μέσω του βαθμού ομοιότητας μεταξύ του πλησιέστερου γείτονα-εγγράφου και του δοκιμαστικού εγγράφου. Αν περισσότεροι του ενός *k*-πλησιέστερων γειτόνων μοιράζονται μία κλάση, τότε προστίθενται οι επιμέρους βαρύτητες, και το άθροισμα αυτών θεωρείται η βαρύτητα αυτής της κλάσης σχέση με το δοκιμαστικό έγγραφο. Ταξινομώντας τις πιθανές κλάσεις με βάση τον εκάστοτε βαθμό βαρύτητάς τους, προκύπτει η απορρέουσα ταξινομημένη λίστα πιθανών κλάσεων για το δοκιμαστικό έγγραφο [80].

Πρόκειται για έναν αλγόριθμο κατηγοριοποίησης που στηρίζεται στη μηχανική μάθηση και αποτελεί ένας από τους κλασικούς αλγορίθμους στο πεδίο της κατηγοριοποίησης κειμένων. Αυτός είναι ένας από τους βασικούς λόγους που στο **ΚΕΦΑΛΑΙΟ 5**: Πειραματική Αξιολόγηση, αξιοποιείται ο *k*-NN για την πραγματοποίηση της συγκριτικής πειραματικής μελέτης που πραγματοποιείται στο πλαίσιο της αξιολόγησης της προτεινόμενης μεθοδολογίας.

3.2.2. Naïve Bayes

Ένας κατηγοριοποιητής *Naïve Bayes* είναι ένας πιθανολογικός κατηγοριοποιητής που βασίζεται στην εφαρμογή τού *Θεωρήματος κατά Bayes*, όπου η τρέχουσα πιθανότητα σχετίζεται με την αρχική και τα χαρακτηριστικά των δεδομένων προς επεξεργασία είναι μεταξύ τους ανεξάρτητα. Τα χαρακτηριστικά αυτά διαφέρουν ανάλογα με το πεδίο εφαρμογής του θεωρήματος. Στην κατηγοριοποίηση κειμένων είναι συνήθως η βαρύτητα όρων (λέξεων) μέσα στο κείμενο, η οποία υπολογίζεται από τον τύπο *tf*idf* ή κάποιο άλλο σχήμα. Αυτοί οι κατηγοριοποιητές αξιοποιούνται συχνά, καθώς είναι γρήγοροι και εύκολα εφαρμόσιμοι [24].

3.2.3. Δέντρα Απόφασης (decision trees)

Τα *Δέντρα Απόφασης* είναι μια μέθοδος κατηγοριοποίησης, η οποία χρησιμοποιεί πρότυπα *διαστήματα διαίρεσης* (instance space division). Πιο συγκεκριμένα, αποτελούνται από *κόμβους* (nodes) και προσανατολισμένες *ακμές* (oriented arcs). Ο *κόμβος ρίζας* (root node) δεν έχει εισερχόμενα *τόξα*. Όλοι οι άλλοι κόμβοι έχουν ακριβώς ένα εισερχόμενο τόξο, αλλά τα εξερχόμενα μπορούν να είναι και περισσότερα του ενός. Τέλος, τα *φύλλα* (leaves) ενός δέντρου απόφασης είναι κόμβοι, οι οποίοι έχουν ένα εισερχόμενο τόξο, αλλά κανένα εξερχόμενο [67]. Κάθε εσωτερικός κόμβος ενός δέντρου απόφασης χωρίζει το πρότυπο διάστημα διαίρεσης σε ένα ή περισσότερα υπο-πρότυπα διαστήματα, σύμφωνα με την εισαχθείσα τιμή γνωρίσματος της διακριτής διαδικασίας. Τα φύλλα ενός δέντρου απόφασης δείχνουν την κατηγοριοποίηση και οι ακμές αναπαριστούν συνδυασμό χαρακτηριστικών.

Εν συντομία, η λειτουργία ενός Δέντρου Απόφασης περιγράφεται ως εξής: αρχικά ένας αλγόριθμος κατηγοριοποίησης και ένα σύνολο δεδομένων εκπαίδευσης χρησιμοποιούνται για τη δημιουργία ενός δέντρου απόφασης, και στη συνέχεια κάθε νέα καταχώριση κατηγοριοποιείται αξιοποιώντας το δέντρο απόφασης. Η νέα καταχώριση προς κατηγοριοποίηση, αφού περάσει από τον κόμβο ρίζας ρου δέντρου, από όπου λαμβάνει την τιμή του αντίστοιχου γνωρίσματος καταχώρισης, οδηγείται σε μια από τις ακμές του δέντρου και βάσει αυτής (δηλαδή της τιμής γνωρίσματος) και συνεχίζει την πορεία της πάνω στις ακμές μέχρι ο κόμβος στον

οποίο θα φτάσει να αποτελεί φύλλο του δέντρου. Τα φύλλα του δέντρου απόφασης αναπαριστούν τις κλάσεις της κατηγοριοποίησης, και έτσι η καταχώριση επισημειώνεται με τον όνομα της κλάσης που φέρει το φύλλο.

Τα Δέντρα Απόφασης έχουν τη δομή ενός δέντρου, όπου κάθε εσωτερικός κόμβος περιέχει έναν όρο, τα κλαδιά που ξεκινούν από κάθε κόμβο περιέχουν το βάρος του όρου, και τα φύλλα τις κατηγορίες. Κατηγοριοποιητές που στηρίζονται στα δέντρα απόφασης κατηγοριοποιούν ένα δοκιμαστικό έγγραφο d_j εξετάζοντας κατ' επανάληψη το βάρος των όρων που περιέχουν οι εσωτερικοί κόμβοι μέχρι να φτάσει σε κάποιο φύλλο. Οι περισσότεροι αξιοποιούν δυαδική αναπαράσταση εγγράφων και γι' αυτό περιλαμβάνουν δέντρα δυαδικών αποφάσεων. Επίσης, μπορούν να αξιοποιηθούν είτε ως βασικό εργαλείο κατηγοριοποίησης είτε σαν βάση για τη δημιουργία κατηγοριοποιητών [74].

3.2.4. Νευρωνικά Δίκτυα (neural networks)

Ένας κατηγοριοποιητής *Νευρωνικού Δικτύου* συνιστά ένα δίκτυο μονάδων, όπου κάθε μονάδα εισόδου αναπαριστά έναν όρο, κάθε μονάδα εξόδου αναπαριστά την/τις κατηγορία/ίες ενδιαφέροντος, και τα βάρη στις ακμές που ενώνουν τις μονάδες αναπαριστούν τις σχέσεις εξάρτησης. Για την κατηγοριοποίηση ενός εγγράφου κειμένου, εισάγονται τα βάρη των όρων του στις μονάδες εισαγωγής, με αποτέλεσμα να «ενεργοποιούνται» οι αντίστοιχοι «νευρώνες» του δικτύου και οι τιμές των μονάδων εξόδου καθορίζουν την απόφαση για την κατηγορία του εγγράφου [74].

3.2.5. Support Vector Machines

Πρόκειται για έναν αλγόριθμο επίλυσης προβλημάτων αναγνώρισης μεταξύ δύο κλάσεων [24]. Ακόμα και στην περίπτωση επίλυσης προβλημάτων πολλαπλών κλάσεων, ο αλγόριθμος αυτός τα μετατρέπει σε επιμέρους προβλήματα αναγνώρισης μεταξύ δύο κλάσεων. Η ακρίβειά του αξιολογείται ως υψηλή, αλλά υψηλές είναι και οι απαιτήσεις του σε υπολογιστικούς πόρους και δυνατότητες [56]. Πρόκειται για έναν από τους πιο διαδεδομένους αλγορίθμους κατηγοριοποίησης εποπτευόμενης

μάθησης, που μπορούν να αποτελέσουν πολύτιμη αφετηρία σχεδιασμού και ανάπτυξης τεχνικών κατηγοριοποίησης σελίδων διαδικτύου [67].

3.3. Τεχνικές Κατηγοριοποίησης Σελίδων Διαδικτύου βάσει Κειμενικής Πληροφορίας

Η αυτοματοποιημένη επεξεργασία, οργάνωση και διαχείριση της διαθέσιμης στο διαδίκτυο κειμενικής πληροφορίας αποτελεί ένα αναμφισβήτητα θεμελιώδες πρόβλημα. Η εξόρυξη γνώσης από κείμενα έχει πολλές σημαντικές εφαρμογές, όπως είναι η ταξινόμηση (δηλαδή, εποπτευόμενη, μη εποπτευόμενη και ημι-εποπτευόμενη ταξινόμηση), το φιλτράρισμα εγγράφων, η συνοπτική παρουσίαση και *ανάλυση συναισθημάτων/ταξινόμηση γνώμης* (opinion classification).

Ένας ταξινομητής κειμένων αναμένεται να ετικετοποιεί έγγραφα κειμένου σε προκαθορισμένες κλάσεις με προφανή παραδοχή ότι κάθε κλάση περιλαμβάνει παρόμοια έγγραφα, μιλώντας συνήθως για ένα συγκεκριμένο θέμα που είναι διαφορετικό από αυτό των άλλων κλάσεων. Ωστόσο, η διανυσματική αναπαράσταση των εγγράφων συνήθως εμφανίζεται ανεπαρκής και συνεπώς με υψηλό βαθμό αδυναμίας. Αυτό συνιστά σημαντικό εμπόδιο, ειδικά όταν υπάρχουν πολλές ετικέτες κλάσεων με ανεπαρκή δεδομένα εκπαίδευσης για καθένα από αυτά. Η απόκτηση ποιοτικά επισημειωμένων δεδομένων για εκπαίδευση είναι συνήθως κοστοβόρα σε εφαρμογές πραγματικού κόσμου. Κατά συνέπεια, ένας ακριβής ταξινομητής κειμένων πρέπει να έχει την ικανότητα να χρησιμοποιεί αυτές τις σημαντικές πληροφορίες της πολυσημίας που συναντάμε στη φυσική γλώσσα.

Μια παραδοσιακή μέθοδος για την αναπαράσταση εγγράφων ονομάζεται ***B**ag **o**f **W**ords* (BOW). Αυτή η τεχνική αναπαράστασης περιλαμβάνει πληροφορίες σχετικές μόνο με τους όρους και τις αντίστοιχες συχνότητές τους σε ένα έγγραφο ανεξάρτητα από τις θέσεις τους μέσα σε μια πρόταση ή σε ένα έγγραφο. Ονομάζεται επίσης ***V**ector **S**pace **M**odel* (VSM) αφού κάθε έγγραφο αντιπροσωπεύεται ως φορέας συχνότητας όρων στο λεξιλόγιο. Επιπλέον, αυτή η αναπαράσταση δεν λαμβάνει υπόψη τις σημασιολογικές συσχετίσεις μεταξύ των λέξεων. Για παράδειγμα, δύο

λέξεις γραμμένες ως διαφορετική ακολουθία χαρακτήρων αποτελούν διαφορετικές ορθογώνιες διαστάσεις αυτού του διανυσματικού χώρου, παρότι μπορεί να είναι συνώνυμες. Επιπλέον, η σειρά αυτών των λέξεων στις προτάσεις χάνεται εντελώς στην αναπαράσταση BOW. Αυτή η προσέγγιση εστιάζει κυρίως στην ύπαρξη κάποιας μορφής πληροφορίας σχετικά με τη συχνότητα των όρων. Η μέθοδος BOW κάνει την αναπαράσταση των εγγράφων απλούστερη αγνοώντας τις συνακόλουθες διαφορετικές σημασιολογικές και συντακτικές σχέσεις μεταξύ λέξεων μιας φυσικής γλώσσας. Πρώτον, αγνοεί τις εκφράσεις που αποτελούνται από πολλές λέξεις και τις διαχωρίζει σε ανεξάρτητους όρους. Δεύτερον, αντιμετωπίζει τις πολύσημες λέξεις (λέξεις με πολλαπλές σημασίες) ως μία, ενιαία οντότητα επειδή η λέξη διαχωρίζεται από τις γειτονικές που καθορίζουν και τη σημασία της. Τρίτον, η προσέγγιση BOW αντιμετωπίζει συνώνυμες λέξεις ως διαφορετικούς όρους [70].

Με τη δημιουργία του σημασιολογικού ιστού, το περιεχόμενο του διαδικτύου άρχισε να αποκτά «νόημα» και για τους υπολογιστές. Στόχος της δημιουργίας του ήταν να παρέχει μεγαλύτερο εύρος λειτουργικότητας μέσω «έξυπνων» εργαλείων. Την ίδια στιγμή, η αναπαράσταση δεδομένων και γνώσης μέσω γλωσσών, π.χ. XML, RDF, κατανοητών στον υπολογιστή, συμβάλλει σημαντικά στη διαχείριση της πληροφορίας διαθέσιμης στο διαδίκτυο, δεδομένου ότι αυξάνεται με ανεξέλεγκτους ρυθμούς. Αποτέλεσμα της παραπάνω εξέλιξης είναι η αξιοποίηση των αλγορίθμων κατηγοριοποίησης για την κατηγοριοποίηση των ιστοσελίδων με βάση την κειμενική τους πληροφορία. Επιπλέον, οι τεχνικές κατηγοριοποίησης διαδικτυακών σελίδων με βάση την κειμενική τους πληροφορία, χρησιμοποιούν τα σημασιολογικά δίκτυα, τις οντολογίες και τις ιεραρχίες για τη δημιουργία ομάδων αντικειμένων, και, αξιοποιώντας τις σχέσεις μεταξύ των κατηγοριών των αντικειμένων, οργανώνουν θεματικά τις σελίδες. Το κοινό σημείο μεταξύ των μεθόδων αυτών είναι ότι εφαρμόζουν τεχνικές προ-επεξεργασίας κειμένων για την εξαγωγή λέξεων κλειδιών, και, στηριζόμενες στην συχνότητας εμφάνισής τους μέσα στο κείμενο και τη σημασιολογική τους εγγύτητα, καταλήγουν στην αντίστοιχη θεματική κατηγορία για κάθε σελίδα προς κατηγοριοποίηση. Το γεγονός αυτό οδηγεί τους επιστήμονες στη συνδυαστική αξιοποίηση των μεθόδων ***N**atural **L**anguage **P**rocessing* (NLP), ***M**achine **L**earning* (ML) και ***D**ata **M**ining* (DM) για τον εντοπισμό προτύπων από τους

διαφορετικούς τύπους των εγγράφων και την ταξινόμησή τους με αυτόματο τρόπο [87].

Γενικότερα, η σημασιολογική κατηγοριοποίηση κειμένων χαρακτηρίζεται από ορισμένα πλεονεκτήματα έναντι της παραδοσιακής κατηγοριοποίησης κειμένων. Στις μεθόδους σημασιολογικής κατηγοριοποίησης κειμένων, λαμβάνονται υπόψη οι σημασιολογικές σχέσεις μεταξύ των λέξεων προκειμένου να υπολογιστεί η ομοιότητα μεταξύ των εγγράφων. Η σημασιολογική προσέγγιση επικεντρώνεται στη σημασία των λέξεων και των κρυφών σημασιολογικών συνδέσεων μεταξύ των λέξεων και κατά συνέπεια μεταξύ των εγγράφων. Τα βασικά πλεονεκτήματα της σημασιολογικής ταξινόμησης κειμένων σε σχέση με την παραδοσιακή ταξινόμηση κειμένων είναι α) η σημασιολογική κατανόηση του κειμένου, που βελτιώνει την ακρίβεια της ταξινόμησης, και β) η ικανότητα χειρισμού συνωνυμίας και πολυσημίας σε σύγκριση με τους παραδοσιακούς αλγορίθμους ταξινόμησης κειμένων, δεδομένου ότι οι εν λόγω τεχνικές αξιοποιούν τις σημασιολογικές σχέσεις μεταξύ λέξεων. Στην παρούσα Ενότητα γίνεται σύντομη αναφορά σε σχετικές προηγούμενες μελέτες, οι οποίες και καθόρισαν σε μεγάλο βαθμό το σχεδιασμό της προτεινόμενης μεθοδολογίας.

3.3.1. Σημασιολογικά Δίκτυα

Εδώ και δεκαετίες έχει γίνει ευρέως αποδεκτή η δομική αναλογία που χαρακτηρίζει τα σημασιολογικά δίκτυα και το διαδίκτυο, με αποτέλεσμα την αξιοποίηση των πρώτων σε μελέτες που αφορούν την οργάνωση των δεδομένων διαδικτύου. Πιο συγκεκριμένα, για τους δημιουργούς δεδομένων διαδικτύου, τα συστήματα υπερκειμένων προσφέρουν μεγάλη ευελιξία στη σύνδεση πληροφοριών, ώστε αυτές να εμφανίζονται σαν μια «συναρμολογημένη» συλλογή στο πλαίσιο ενός πληροφοριακού δικτύου. Έτσι, τα υπερ-κείμενα μπορούν να αποτελέσουν ένα πολύτιμο μέσο οργάνωσης της πληροφορίας, με σκοπό την εξυπηρέτηση των αναγκών των χρηστών. Την ίδια στιγμή, για τους χρήστες, τα συστήματα υπερ-κειμένων παρέχουν εργαλεία πλοήγησης σε αυτά τα πληροφοριακά δίκτυα. Από την άλλη πλευρά, τα σημασιολογικά δίκτυα, συνιστούν έναν ιδιαιτέρως οργανωμένο γλωσσικό πόρο, ο οποίος δίνει τη δυνατότητα ευέλικτης πλοήγησης μεταξύ των

λεξικών αντικειμένων μιας γλώσσας. Η δομική αναλογία, λοιπόν, μεταξύ του διαδικτύου και των σημασιολογικών δικτύων είναι ότι τα τελευταία αποτελούν ένα είδος σχήματος αναπαράστασης γνώσης, το οποίο αποτελείται από έναν κατευθυνόμενο γράφο, όπου οι εννοιολογικές μονάδες εμφανίζονται ως κόμβοι και οι μεταξύ τους σχέσεις ως σύνδεσμοι. Ο γράφος αποκτά σημασιολογικό χαρακτήρα όταν κάθε κόμβος και κάθε σύνδεσμος επισημειώνονται αποκτώντας νόημα, π.χ. ιεραρχική σχέση, σχέση συνωνυμίας/αντωνυμίας κ.ο.κ. [8].

Δεδομένων των παραπάνω, εδώ και δεκαετίες έχει αποδειχθεί πως η εννοιολογική αποσαφήνιση βελτιώνει σημαντικά τη διαδικασία της ανάκτησης πληροφορίας. Πιο συγκεκριμένα, ακόμα και στην περίπτωση που η εννοιολογική αποσαφήνιση μέσω των σημασιολογικών δικτύων δεν είναι άριστη, η κατηγοριοποίηση που προκύπτει αξιοποιώντας τα συμβάλλει σημαντικά στην αποτελεσματικότερη ανάκτηση πληροφορίας συγκριτικά με την κατηγοριοποίηση αξιοποιώντας μοντέλα με όρους βαρύτητας [31]. Ακόμα ένα πεδίο, όπου βρίσκει εφαρμογή η αξιοποίηση των λεξικογραφικών οντολογιών για την κατηγοριοποίηση εγγράφων, είναι η ικανοποίηση *διαγλωσσικών πληροφοριακών αιτημάτων* (cross-language queries). Σε αυτές τις περιπτώσεις, αξιοποιούνται οι σχέσεις υπωνυμίας προκειμένου το ερώτημα να επεκταθεί και να βελτιωθεί η διαδικασία της αποσαφήνισης [1].

Το *WordNet* είναι ένα είδος τυπικού λεξικού, το οποίο περιλαμβάνει και τη σημασιολογία των όρων που περιέχει. Οι κύριες δομικές οντότητες του WordNet είναι οι εσωτερικές γλωσσολογικές σχέσεις, μέσω των οποίων οι λέξεις συνδέονται βάσει των σημασιολογικών τους σχέσεων. Η κύρια συμβολή του WordNet στη λεξικογραφία είναι τα συστηματικά μοτίβα και οι συστηματικές σχέσεις που υπάρχουν μεταξύ των εννοιών και εκφράζονται μέσω λεξιλογικών μονάδων. Από αυτή τη σκοπιά, το WordNet, ως ιδιαίτερος τύπος σημασιολογικού δικτύου, μοιάζει περισσότερο με υπερκείμενο, όσον αφορά τη δομική οργάνωση της πληροφορίας που περιέχει. Επιλέγουμε να αναφερθούμε σε αυτό, αφού εκτός από την ευρέως αποδεκτή σημασία που έχει, αποτελεί και ένα από τα εργαλεία που αξιοποιούμε στο πλαίσιο της προτεινόμενης μεθοδολογίας.

3.3.2. Οντολογίες

Για την Επιστήμη της Πληροφορίας, μια οντολογία αποτελεί ένα σχήμα αναπαράστασης γνώσης, αποτελούμενο από το όνομα και τον ορισμό κάθε κατηγορίας, τις ιδιότητες της καθεμιάς και τις μεταξύ τους σχέσεις. Με άλλα λόγια, οι οντολογίες αναπαριστούν τις ιδιότητες μιας θεματικής περιοχής και πώς αυτές σχετίζονται, μέσω του ορισμού ενός συνόλου εννοιών και κατηγοριών που αναπαριστούν την κάθε θεματική περιοχή. Ένας από τους πρώτους ορισμούς της οντολογίας δόθηκε από τον Neches [54] ο οποίος όρισε την οντολογία ως εξής: *μία οντολογία ορίζει τους βασικούς όρους και τις σχέσεις που περιλαμβάνει το λεξιλόγιο μιας θεματικής περιοχής, καθώς και τους κανόνες για το συνδυασμό όρων και σχέσεων προκειμένου να οριστούν οι επεκτάσεις του λεξιλογίου.* Μερικά χρόνια αργότερα, ο Gruber [32] όρισε την οντολογία ως μια *ρητή εξειδίκευση της σύλληψης μιας έννοιας* (explicit specification of a conceptualization).

Οι οντολογίες αξιοποιούνται στο πλαίσιο ποικίλων επιστημονικών πεδίων, όπως είναι το ηλεκτρονικό εμπόριο, η διαχείριση γνώσης, η επεξεργασία φυσικής γλώσσας και άλλα, καθώς αποτελούν στοιχείο-κλειδί για την αντιμετώπιση της σημασιολογικής ετερογένειας. Παράλληλα, μέσω των οντολογιών επιτυγχάνεται η «επικοινωνία» μεταξύ ανθρώπου και υπολογιστικών εφαρμογών, καθώς αποτελούν ένα σχήμα οργάνωσης της γνώσης κατανοητό εκατέρωθεν [16]. Από τη σκοπιά της ανάκτησης πληροφορίας, ο συνδυασμός σημασιολογίας και στατιστικής αυξάνει τις πιθανότητες ανάκτησης ακριβέστερων εγγράφων και γλωσσολογικά σχετικότερων με τις προς αναζήτηση πληροφορίες. Ωστόσο, στο πλαίσιο του διαδικτύου η διαδικασία αυτή δυσχεραίνεται, καθώς δεν είναι όλες οι σελίδες επισημειωμένες σημασιολογικά. Για το λόγο αυτό, οι [22] στηρίζονται στη σημασιολογική επισημείωση των σελίδων αξιοποιώντας εξωτερικούς πόρους (γλωσσολογικές οντολογίες) για τη σημασιολογική αποσαφήνιση του περιεχομένου τους.

Σε άλλη προσέγγιση κατηγοριοποίησης εγγράφων με τη χρήση οντολογιών, συμπεριλαμβάνονται οι τεχνικές εκμετάλλευσης κειμένων και η κατασκευή οντολογίας [84]. Πιο συγκεκριμένα, χρησιμοποιούν το εργαλείο OntoGen (εργαλείο

για την ημι-αυτόματη δημιουργία οντολογιών) για την αυτόματη κατασκευή οντολογίας και για την ανάθεση κάθε άρθρου στην κατηγορία που ανήκει. Στη δουλειά των [85], αξιοποιείται επίσης η οντολογία προκειμένου να μειωθούν οι λεξιλογικές αμφισημίες κατά την κατηγοριοποίηση εγγράφων. Επιπρόσθετα στα παραπάνω, οι οντολογίες έχουν αξιοποιηθεί εξίσου στην κατηγοριοποίηση της ηλεκτρονικής αλληλογραφίας [79] και των ηλεκτρονικών εφημερίδων [81]. Δεδομένου ότι η χειρωνακτική ετικετοποίηση απαιτεί αξιοσημείωτη ανθρώπινη προσπάθεια, πολλοί ερευνητές αξιοποιούν υποσύνολα υπάρχοντων διαδικτυακών ευρετηρίων, τα οποία είναι στη διάθεση του κοινού [64].

Στη μελέτη [69], εισάγεται ένα νέο πλαίσιο για την κατηγοριοποίηση σελίδων διαδικτύου, το οποίο κατηγοριοποιεί κάθε σελίδα σε μία από τις προκαθορισμένες κλάσεις. Η μεθοδολογία αυτή ονομάζεται ***C****lassification using* ***M****ulti-layered* ***D****omain* ***O****ntology* (CMDO), αξιοποιεί διάφορες τεχνικές εξόρυξης γνώσης από κείμενα και απαρτίζεται από τρία επιμέρους μοντέλα: i) *Page Analysis Module* (PAM), το οποίο αναλύει την υπό επεξεργασία σελίδα λαμβάνοντας υπόψη τη δομή της, ii) *Page Importance Module* (PIM), το οποίο αντιπροσωπεύει έναν «διυλιστή» που αφήνει να «περάσουν» μόνο οι σελίδες που σχετίζονται με τον τομέα ενδιαφέροντος, iii) *Page Classification Module* (PCM), το οποίο κατηγοριοποιεί τη σελίδα στις προκαθορισμένες κλάσεις. *Multi-Layered Domain Ontology* (MLDO) είναι ένας εννοιολογικό γράφος, όπου κάθε έννοια είναι επισημειωμένη με το «βάρος» της, και ο οποίος συνδυάζει πολλά επίπεδα. Κάθε επίπεδο είναι μια οντολογία που αντιπροσωπεύει μία από τις κλάση του τομέα. Τα πειραματικά αποτελέσματα έδειξαν ότι το προτεινόμενο πλαίσιο κατηγοριοποίησης ξεπερνά τα υπάρχοντα, καθώς κατηγοριοποιεί με μεγαλύτερη ακρίβεια και ανάκληση.

3.3.2.1. Η Wikipedia ως οντολογία

Η ηλεκτρονική εγκυκλοπαίδεια *Wikipedia*, ως ένα κοινωνικό φαινόμενο συνεργατικής δημιουργίας γνώσης, έχει μελετηθεί εκτενώς από διάφορες σκοπιές. Αυτό που είναι αδιαμφισβήτητο, όμως, είναι πως πρόκειται για την πιο συχνά επισκέψιμη ελεύθερη πηγή πληροφοριών του διαδικτύου. Κάνοντας μια ιστορική αναδρομή στην

οργάνωση των πληροφοριών που περιέχει, τα πρώτα χρόνια μετά την δημιουργία της, 2001-2004, δεν έχει κάποιο σύστημα για την οργάνωση των άρθρων της. Η μόνη έκφραση δομικής οργάνωσης είναι οι σύνδεσμοι στο σώμα ενός άρθρου που οδηγεί απευθείας σε κάποιο άλλο άρθρο, χωρίς να υπάρχει κάποιο εργαλείο περαιτέρω οργάνωσης. Από το 2004, η Wikipedia προσθέτει ένα χαρακτηριστικό ειδικά σχεδιασμένο για αυτό το σκοπό. Δημιουργούνται σελίδες με κατηγορίες, οι οποίες αποτελούν κυρίως συλλογές συνδέσμων που οδηγούν σε άρθρα ή άλλες σελίδες με κατηγορίες [78]. Σταδιακά, αυτό επεκτείνεται και πλέον η Wikipedia χαρακτηρίζεται από ένα ολοκληρωμένο σύστημα κατηγοριών με ιεραρχική δομή, μέσα από το οποίο οι χρήστες – γνωρίζοντας τα χαρακτηριστικά που ορίζουν ένα θέμα – μπορούν γρήγορα και εύκολα να εντοπίσουν και να πλοηγηθούν σε ένα σύνολο σελίδων με τα ίδια χαρακτηριστικά [II]. Έτσι, προοδευτικά η Wikipedia έχει αποδειχθεί και αναγνωριστεί και ως οντολογία με ιεραρχική δομή, συνιστώντας ένα εργαλείο χρήσιμο στον χώρο της κατηγοριοποίησης δεδομένων, ενώ σε άλλες περιπτώσεις αξιοποιείται για την επέκταση κατηγοριοποιήσεων με ιεραρχική δομή, όπως στην περίπτωση των [34].

Ενδεικτικά, στην δουλειά των [52] παρουσιάζεται μια προσέγγιση δημιουργίας σημασιολογικού δικτύου στηριζόμενου στη Wikipedia, κάνοντας να διαλειτουργήσουν τεχνικές εξαγωγής πληροφοριών από ανοιχτές πηγές με τεχνικές απόκτησης γνώσης. Ο αλγόριθμος που προτείνουν, εξάγει παραδείγματα σχέσεων από το σώμα σελίδων της Wikipedia και τους δίνει χαρακτηριστικά οντολογίας α) δημιουργώντας σύνολα συνώνυμων σχεσιακών φράσεων, δηλαδή σημασιολογικά ισοδύναμα σύνολα φράσεων, β) αναθέτοντας σημασιολογικές κλάσεις στα σχεσιακά αυτά σημασιολογικά ισοδύναμα σύνολα και γ) αποσαφηνίζοντας τα αρχικά παραδείγματα σχέσης με σχέσεις σημασιολογικά ισοδύναμων φράσεων. Αυτό έχει σαν αποτέλεσμα την δημιουργία τού WiSeNet, ενός σημασιολογικού δικτύου όπου οι σελίδες της Wikipedia αποτελούν έννοιες και ετικέτες, που μεταξύ τους έχουν σχέσεις οντολογίας.

Σε άλλη δουλειά, [49], προκειμένου να αντιμετωπιστεί το πρόβλημα του εντοπισμού θεμάτων που σχετίζονται με ένα σύνολο εγγράφων, προτείνεται ένα εργαλείο, το

TagTheWeb, ως μια γενική μέθοδος κατηγοριοποίησης που βασίζεται στη γνώση που εκφράζεται από την ταξινομική δομή της Wikipedia. Αυτό στηρίζεται στην παραγωγή δακτυλικών αποτυπωμάτων μέσω της σημασιολογικής σχέσης μεταξύ κόμβων του γραφήματος κατηγοριών της Wikipedia, και μπορεί να χρησιμοποιηθεί – σ.σ. το TagTheWeb - ως διαδικτυακή διεπαφή ή διαδικτυακή διεπαφή προγραμματισμού εφαρμογών για την ταξινόμηση οποιουδήποτε πόρου βάσει κειμένου.

Οι [3] προτείνουν ένα πιθανολογικό μοντέλο θεμάτων, που ενσωματώνει τη γνώση της DBpedia για την ανάθεση ετικετών σε σελίδες διαδικτύου και άλλα διαδικτυακά έγγραφα με βάση τα θέματα που πραγματεύονται. Η μεθοδολογία τους βασίζεται στην ολοκλήρωση του δικτύου ιεραρχικών κατηγοριών DBpedia με μοντέλα στατιστικών θεμάτων, όπου οι κατηγορίες DBpedia θεωρούνται θέματα. Ενδεικτική είναι και η δουλειά των [73], που χρησιμοποιούν τίτλους και κατηγορίες άρθρων της Wikipedia για να προσδιορίσουν θέματα εγγράφων. Σύμφωνα με τη μεθοδολογία τους, αρχικώς εντοπίζονται όλα τα σχετικά άρθρα της Wikipedia σε ένα έγγραφο ταιριάζοντας τους τίτλους τους με τις λέξεις του εγγράφου, και στη συνέχεια επιλέγονται οι κατηγορίες που αντιστοιχούν σε αυτά τα άρθρα, ταξινομούνται, και τελικά επιλέγονται οι κατηγορίες με τα υψηλότερα βάρη ως τα θέματα του εγγράφου.

Στην περίπτωση [37] η ίδια η οντολογία, συμπεριλαμβανομένων των εννοιών σχετικών με έναν τομέα που οργανώνονται σε ιεραρχίες κατηγοριών και διασυνδέονται μέσω σχέσεων, καθώς και των παραδειγμάτων σύνδεσης μεταξύ τους, αποτελεί τον ταξινομητή. Η μεθοδολογία εστιάζει (i) στην μετατροπή ενός εγγράφου κειμένου σε θεματικό γράφημα οντοτήτων που εμφανίζονται στο έγγραφο, (ii) στην οντολογική κατηγοριοποίηση των οντοτήτων στο γράφημα και (iii) στον προσδιορισμό της συνολικής κατηγοριοποίησης του θεματικού γραφήματος, και κατά συνέπεια του ίδιου του εγγράφου. Στα πειράματα που παρουσιάζουν χρησιμοποιούν μία οντολογία RDF που κατασκευάστηκε από την πλήρη αγγλική έκδοση της Wikipedia.

Τέλος, την ιεραρχική δομή της Wikipedia έχουμε μελετήσει και σε προηγούμενη δουλειά μας [55], όπου προτείνεται και υλοποιείται ένα πρότυπο μοντέλο ιεραρχικής οργάνωσης δεδομένων, το οποίο κάνοντας εκτεταμένη χρήση ενός σημασιολογικού δικτύου λημμάτων, εντοπίζει εννοιολογικές σχέσεις μεταξύ των θεματικών κατηγοριών υπό τις οποίες οργανώνονται τα άρθρα της Wikipedia και με βάση αυτές επιχειρεί και κατορθώνει την αυτοματοποιημένη και αποδοτική ιεραρχική οργάνωση των άρθρων. Με άλλα λόγια, εξετάζεται η συνεισφορά των σημασιολογικών δικτύων στην επεξεργασία και αποσαφήνιση των ονομάτων των θεματικών κατηγοριών της Βικιπαίδεια, στη λημματοποίησή τους και εν τέλει στην ιεραρχική τους οργάνωση. Στηριζόμενοι σε αυτή τη δουλειά, αξιοποιούμε τη Wikipedia και στο πλαίσιο της παρούσας διδακτορικής διατριβής, όπως φαίνεται στην αναλυτική περιγραφή της προτεινόμενης μεθοδολογίας (**4.3.** Αναλυτική Περιγραφή Μεθοδολογίας).

3.3.3. Ιεραρχίες

Υπάρχουν αρκετές δημοσιευμένες προσεγγίσεις για την ιεραρχική κατηγοριοποίηση εγγράφων βάσει κειμενικής πληροφορίας όπου αξιοποιούνται οι ιεραρχίες. Μεταξύ αυτών, παρότι όχι αποδεσμευμένη από την ανθρώπινη παρέμβαση, ξεχωρίζει η προσέγγιση των [44], όπου αξιοποιούνται οι κατηγορίες *Yahoo!-Categories* ως μια εννοιολογική ιεραρχία με σκοπό την κατηγοριοποίηση εγγράφων. Στην ίδια κατεύθυνση, αλλά για την θεματική κατηγοριοποίηση σελίδων διαδικτύου αξιοποιούνται οι *διαδικτυακοί κατάλογοι* (web directories) ως θεματικές ιεραρχίες [2][77]. Σε άλλες μελέτες οι κατηγορίες της εγκυκλοπαίδειας Wikipedia αξιοποιούνται επίσης σαν πηγή ιεραρχικής δομής για την κατηγοριοποίηση διαδικτυακών εγγράφων [21]. Ομοίως και το σημασιολογικό δίκτυο WordNet που αναφέρεται σε προηγούμενη Ενότητα (**3.3.1.** Σημασιολογικά Δίκτυα).

Στη μελέτη των [60] παρουσιάζεται ένας αυτόματος αλγόριθμος ταξινόμησης σελίδων διαδικτύου. Ο αλγόριθμος αυτός, σε αντίθεση με άλλους, τονίζει το ζήτημα της δυναμικής ανάπτυξης του παγκόσμιου ιστού και λαμβάνει υπόψη τις πληροφορίες ιεραρχικής δομής για τη βελτίωση της ακρίβειας της κατηγοριοποίησης. Ο πυρήνας του αλγορίθμου είναι μία τεχνική ιεραρχικής κατηγοριοποίησης που

αποδίδει μια ιστοσελίδα σε μια κατηγορία. Ο αριθμός των σελίδων διαδικτύου στον Ιστό αυξάνεται συνεχώς με μεγάλη ταχύτητα και επομένως είναι αδύνατο για ένα καθορισμένο σύνολο κατηγοριών να παρέχεται ακριβής κατηγοριοποίηση. Για την αντιμετώπιση αυτού του προβλήματος, προτείνουν και εφαρμόζουν μια δυναμική τεχνική επέκτασης.

3.4. Τεχνικές Κατηγοριοποίησης Σελίδων Διαδικτύου βάσει Δομικών Χαρακτηριστικών

Μελετώντας τη βιβλιογραφία γύρω από το ζήτημα της κατηγοριοποίησης διαδικτυακών σελίδων, εύκολα διαπιστώνει κάποιος πως η κατηγοριοποίησή τους βάσει δομής έχει απασχολήσει λιγότερο την ερευνητική κοινότητα. Αυτό προκύπτει από την ποικιλομορφία που χαρακτηρίζει τη δομή των σελίδων αυτών, αλλά και από το γεγονός ότι είναι πιο «φτωχή» σε πληροφορία. Επίσης, οι περισσότερες ερευνητικές μελέτες ασχολούνται κυρίως με την εμφάνιση που έχουν οι σελίδες μέσω του περιηγητή και λιγότερο με την εμφάνισή τους από την πλευρά της σύνταξης και δημιουργίας τους. Σε ένα βαθμό, αυτό μπορεί να εξηγηθεί μέσω του γεγονότος ότι ενίοτε, ενώ μπορεί να αλλάξουν κάποιες από τις *html ετικέτες* (html tags) μιας σελίδας, η εμφάνισή της παραμένει ίδια. Σε κάθε περίπτωση, οι σελίδες διαδικτύου, γραμμένες σε γλώσσα HTML, περιέχουν αρκετές πρόσθετες πληροφορίες μέσω των html tags, των υπερ-συνδέσμων και του κειμένου που εμφανίζεται και πρέπει να επιλεγεί για να ενεργοποιηθεί ο σύνδεσμος και να οδηγήσει σε μια άλλη ιστοσελίδα (anchor text). Οι πληροφορίες αυτές μπορεί να περιέχονται σε δύο ειδών στοιχεία: αυτά που βρίσκονται στην ίδια τη σελίδα και εκείνα που βρίσκονται στους «γείτονες» της σελίδας αυτής, δηλαδή στις σελίδες που με κάποιον τρόπο σχετίζονται με την σελίδα προς κατηγοριοποίηση [63] Επίσης, πληροφορίες μπορούν να εκμαιευτούν και από το URL των σελίδων, γεγονός που μπορεί να απαλλάξει τη διαδικασία της κατηγοριοποίησης από το «κατέβασμα» της σελίδας, μειώνοντας έτσι τον απαιτούμενο χρόνο και τον απαραίτητο αποθηκευτικό χώρο.

3.4.1. HTML Elements

Αξιοποιώντας τη δομή της HTML [18] για την αναπαράσταση των σελίδων διαδικτύου, μπορούμε να επιλέξουμε το πώς ένας όρος είναι αντιπροσωπευτικός της σελίδας, λαμβάνοντας υπόψη το στοιχείο HTML στο οποίο υπάρχει. Για παράδειγμα, μπορούμε να αναπαραστήσουμε μια σελίδα διαδικτύου χρησιμοποιώντας μόνο τις λέξεις του τίτλου, δηλαδή τις λέξεις που εξάγονται από το στοιχείο του τίτλου (TITLE). Για να πετύχουμε καλή απόδοση στην αναπαράσταση ιστοσελίδων αξιοποιώντας τη δομή HTML, είναι σημαντικό να γνωρίζουμε πού μπορούν να βρεθούν οι πιο αντιπροσωπευτικές λέξεις. Για παράδειγμα, μπορούμε να πιστεύουμε ότι μια λέξη που υπάρχει στο στοιχείο του τίτλου (TITLE) είναι γενικά πιο αντιπροσωπευτική του περιεχομένου του εγγράφου από μια λέξη που υπάρχει στο στοιχείο που ορίζει το κυρίως σώμα κειμένου της σελίδας (BODY) [66]. Με άλλα λόγια, τα HTML tags επιτρέπουν τον ορισμό των στοιχείων που με τη σειρά τους καθορίζουν τη μορφή και το περιεχόμενο των σελίδων διαδικτύου και μπορούν να συνεισφέρουν σημαντικά στην κατηγοριοποίηση των τελευταίων. Μέσα από τη δουλειά των [29] φαίνεται η σημασία των ετικετών αυτών και έχει αποδειχθεί πως υπάρχουν τέσσερα στοιχεία που έχουν ξεχωριστή σημασία, ο τίτλος, οι επικεφαλίδες, τα μεταδεδομένα και το κυρίως κείμενο της σελίδας, και πως όταν αυτά συνδυαστούν, δίνουν το καλύτερο δυνατό αποτέλεσμα κατηγοριοποίησης.

Σε άλλη δουλειά, [61], υποστηρίζεται πως οι μετα-ετικέτες (metatags) ενίοτε αποδεικνύονται καταλληλότερη πηγή στοιχείων για την κατηγοριοποίηση των σελίδων, σπάνια έχουν πλεονάζουσα πληροφορία και έχουν ξεκάθαρο νόημα χαρακτηριστικά που οπωσδήποτε διευκολύνουν τη διαδικασία της κατηγοριοποίησης. Έτσι, στη μελέτη των [61], προτείνεται η κατηγοριοποίηση διαδικτυακών ιστοτόπων αξιοποιώντας τις HTML ετικέτες.

Η μελέτη των [41] πραγματεύεται την απόδοση ενός συστήματος κατηγοριοποίησης σελίδων διαδικτύου που στηρίζεται σε URL. Για το σκοπό αυτό, αξιοποιούνται τα εξής στοιχεία: i) URL - τα κειμενικά τμήματα URL της σελίδας, ii) Anchor text - όλα τα διακριτικά του κειμένου που περιέχονται σε άγκυρες που οδηγούν στη σελίδα iii) Title

text - όλα τα διακριτικά στην ετικέτα τίτλου της ιστοσελίδας. iv) Text - όλα τα διακριτικά του σώματος κειμένου του εγγράφου προέλευσης, εξαιρουμένων των ετικετών στην κεφαλή. Παρότι τα περισσότερα html στοιχεία μπορεί να χρησιμοποιούνται με διαφορετικό τρόπο από τους σχεδιαστές των σελίδων, αποδεικνύεται πως είναι συμβάλλουν σημαντικά στην ακριβέστερη κατηγοριοποίηση των σελίδων.

3.4.2. URL

Κάθε σελίδα διαδικτύου χαρακτηρίζεται από το **U**niversal **R**esource **L**ocator **(URL)**, μια σειρά από χαρακτήρες που ορίζει το «μονοπάτι» που οδηγεί στη σελίδα. Εκτός από τις δυνατότητες των ετικετών HTML, μια ιστοσελίδα μπορεί να κατηγοριοποιηθεί με βάση τη δική της διεύθυνση URL. Τα URL συμβάλλουν σημαντικά στην κατηγοριοποίηση των σελίδων για διάφορους λόγους. Πρώτον, μία διεύθυνση URL είναι εύκολα ανακτήσιμη, κάθε διεύθυνση URL αφορά αποκλειστικά μία σελίδα και κάθε σελίδα έχει μία και μοναδική διεύθυνση URL. Δεύτερον, όταν για την κατηγοριοποίηση των σελίδων στηριζόμαστε αποκλειστικά στη διεύθυνση URL, δεν απαιτείται η λήψη ολόκληρης της σελίδας. Έτσι, τείνει να είναι η κατάλληλη τεχνική για την κατηγοριοποίηση σελίδων που δεν υπάρχουν πια ή που η λήψη τους είναι αδύνατη, και ιδίως στις περιπτώσεις που ο χρόνος ή/και ο χώρος είναι κρίσιμος (π.χ. κατηγοριοποίηση σε πραγματικό χρόνο) [71].

Είναι χαρακτηριστικές οι δουλειές των [42] και των [41] όπου αποδεικνύεται πως οι σελίδες διαδικτύου μπορούν να κατηγοριοποιηθούν με βάση τα URLs τους. Μια τέτοια προσέγγιση, παρότι δίνει αποτελέσματα που σε σημεία υστερούν σε ακρίβεια, ενισχύει την απαλλαγή της διαδικασίας της κατηγοριοποίησης από το «κατέβασμα» και την αποθήκευση των σελίδων, και παράλληλα δίνει λύση στις περιπτώσεις όπου οι ιστοσελίδες δεν περιλαμβάνουν αρκετό κείμενο.

Στην προσέγγιση των [42] επίσης τεκμηριώνεται η χρησιμότητα του URL μόνο στην κατηγοριοποίηση των σελίδων διαδικτύου. Αυτή η τεχνική, όπως υποστηρίζεται και σε αυτή τη μελέτη, αποδεικνύεται ταχύτερη από τις παραδοσιακές τεχνικές κατηγοριοποίησης σελίδων διαδικτύου, καθώς οι τελευταίες δεν χρειάζεται να

ανακτηθούν και να αναλυθούν. Τα αποτελέσματα της έρευνάς τους δείχνουν ότι, σε ορισμένες περιπτώσεις, οι μέθοδοι που βασίζονται στη διεύθυνση URL των σελίδων προσεγγίζουν την απόδοση των σύγχρονών τους προηγμένων μεθόδων κατηγοριοποίησης πλήρους κειμένου, αλλά και αυτών που βασίζονται σε συνδέσμους. Επίσης, αξίζει να αναφερθεί πως η διεύθυνση URL μιας ιστοσελίδας είναι η πιο «οικονομική» να αποκτηθεί μεταξύ των πηγών που φέρουν πληροφορίες χρήσιμες για την κατηγοριοποίηση. Επίσης, η διεύθυνση URL είναι ένα «λακωνικό» στοιχείο, χωρίς περίσσεια πληροφορία. Είναι και καθολικό στοιχείο αφού όλες οι ιστοσελίδες, ανεξάρτητα από το αν είναι προσβάσιμες ή όχι, έχουν URL διεύθυνση. Ακόμα, τα URL's, που σηματοδοτούν τη διεύθυνση μιας σελίδας στο διαδίκτυο, είναι συχνά αναγνώσιμα και κατανοητά από τον άνθρωπο, ενώ μπορούν να δηλώνου έμμεσα την κατηγορία του πόρου [41].

Σε άλλη δουλειά, [5], όπου αξιοποιούνται οι δομικές ιδιότητες των ιστοτόπων, αποδεικνύεται πως υπάρχει έντονη συσχέτιση μεταξύ της δομής ενός συνδέσμου και της λειτουργικότητάς του, του ρόλου του στο διαδίκτυο. Έτσι, στο πλαίσιο της δουλειάς τους, στηρίζονται στις δομικές πληροφορίες ενός ιστοτόπου, για να προσδιορίσουν τη λειτουργικότητά του, π.χ. μηχανή αναζήτησης, κατάλογος διαδικτύου, εταιρικός ιστότοπος κ.ο.κ. Στην ίδια κατεύθυνση, οι [46] επεξεργάζονται περεταίρω τη σχέση μεταξύ της δομής ενός ιστοτόπου και της λειτουργικότητάς τους, γεγονός που αξιοποιούμε και στην προτεινόμενη μεθοδολογία (**4.3.1.** Πολυδιάστατη Κατηγοριοποίηση Σελίδων Διαδικτύου).

3.4.3. (Υπερ-) Σύνδεσμοι

Ήδη δύο δεκαετίες πριν, στη δουλειά των [25] προτείνεται η αναπαράσταση των σελίδων μέσω των συνδέσμων που οδηγούν σε εκείνες και όχι μέσω εκείνων που περιέχουν και στους οποίους οδηγούν οι σελίδες. Η προσέγγιση αυτή αύξησε την ακρίβεια της κατηγοριοποίησης. Ωστόσο, πρόκειται για τεχνική που είναι δύσκολο να εφαρμοστεί, καθώς είναι δύσκολο να εντοπιστούν σύνολα σελίδα που να οδηγούν στη σελίδα υπό επεξεργασία. Επιπλέον, οι τεχνικές που αξιοποιούν τους συνδέσμους που οδηγούν στη σελίδα, όπως και εκείνες που αξιοποιούν τους περιεχόμενους σε

αυτή συνδέσμους που θα δούμε στη συνέχεια, προϋποθέτουν την αποθήκευση των σελίδων, με αποτέλεσμα να απαιτούν υπολογιστικό χώρο και χρόνο.

Στη δουλειά [66] γίνεται προσπάθεια να αναπτυχθεί ένα νέο είδος αναπαράστασης για συνδεδεμένες σελίδες. που χρησιμοποιεί τοπικές πληροφορίες, έτσι ώστε η υπερκειμενική φύση των ιστοσελίδων να μπορεί να αξιοποιηθεί ακόμη και σε περιπτώσεις κατηγοριοποίησης σε πραγματικό χρόνο. Η ιδέα είναι να αξιοποιηθεί η δομή HTML για την αναπαράσταση των σελίδων χωρίς να χρειάζεται να αποθηκευτούν. Για παράδειγμα, αν σε μία σελίδα έχουμε το στοιχείο Α, «*PHP tutorial»*, το στοιχείο Α χρησιμοποιείται για τη σύνδεση της τρέχουσας σελίδας με μία άλλη σελίδα. Οι χρήστες μπορούν να καταλάβουν για τι μιλάει η συνδεδεμένη σελίδα μέσω του περιεχομένου του στοιχείου Α, δηλαδή του «*PHP tutorial*». Όταν ένας προγραμματιστής συνδέει μία σελίδα με άλλη-ες, εισάγει με λίγα λόγια το περιεχόμενο της συνδεδεμένης σελίδας. Μπορούμε εύκολα να υποθέσουμε ότι οι λέξεις που χρησιμοποιούνται σε αυτήν την περιγραφή είναι κοντά στο θέμα της συνδεδεμένης σελίδας και, στη συνέχεια, μπορούμε να χρησιμοποιήσουμε αυτήν την περιγραφή για την αναπαράσταση της συνδεδεμένης σελίδας χωρίς να χρειαστεί να την αποθηκεύσουμε.

Οι [83] προτείνουν διαφορετικούς τύπους χαρακτηριστικών που μπορούν να εξαχθούν από HTML έγγραφα, και αξιολογούν τη χρησιμότητά τους στην κατηγοριοποίηση υπερ-συνδέσμων. Τα αποτελέσματα που πήραν έδειξαν πως στοιχεία, όπως είναι το οι λέξεις που επιλέγοντάς τις οδηγούν στο άνοιγμα της σελίδας (anchor text), η παράγραφος που περιέχει το σύνδεσμο, μπορούν να βελτιώσουν σημαντικά τη διαδικασία της κατηγοριοποίησης, ενώ άλλα, όπως οι επικεφαλίδες πριν το σύνδεσμο, μπορεί να είναι πιο χρήσιμες αν συνδυαστούν με στοιχεία.

Μία ακόμα πηγή πληροφορίας εκτός από το περιεχόμενό τους γραμμένο σε φυσική γλώσσα, είναι οι σύνδεσμοι που περιέχονται σε αυτές. Ιδίως όταν στο περιεχόμενο της σελίδας οι κειμενικές πληροφορίες είναι περιορισμένες ή απουσιάζουν, οι σύνδεσμοι μπορούν να φανούν εξαιρετικά χρήσιμοι στους κατηγοριοποιητές. Αυτή η

τεχνική ανήκει στην ευρύτερη προσέγγιση αξιοποίησης των «γειτονικών» σελίδων μιας σελίδας διαδικτύου για την κατηγοριοποίησή της, όταν η ίδια δεν περιέχει αρκετή πληροφορία [64], και οι υπερ-σύνδεσμοι είναι ο προφανέστερος τρόπος σύνδεσης με άλλες σελίδες.

Μιλώντας για σελίδες-γείτονες και τη χρησιμότητά τους κατά την κατηγοριοποίηση μιας σελίδας διαδικτύου, οι [62] καταδεικνύουν πως οι σελίδες-αδέρφια είναι πολύ χρήσιμες για αυτό το σκοπό. Ενώ, εκείνες που έχουν ρόλο γονέα ή παιδιού σε σχέση με τη σελίδα προς κατηγοριοποίηση, ίσως δημιουργούν και «θόρυβο». Προκειμένου να επιτευχθεί ακριβέστερη κατηγοριοποίηση σε σελίδες που έχουν τα ίδια χαρακτηριστικά, χρησιμοποιούνται οι σελίδες σε ρόλο «αδελφού», καθώς φαίνεται ότι αυτές οι σελίδες έχουν μεγαλύτερη ομοιότητα όταν αφορούν το ίδιο θέμα, και την ίδια στιγμή είναι περισσότερο διαθέσιμες προς χρήση από άλλες [72].

ΚΕΦΑΛΑΙΟ 4: Προτεινόμενη Μεθοδολογία

4.1. Εισαγωγή

Στο πλαίσιο της παρούσας διατριβής, στόχος είναι η διαχείριση και η οργάνωση των δυναμικών δεδομένων, και πιο συγκεκριμένα η συνδυαστική κατηγοριοποίησή τους μέσω της δομικής και σημασιολογικής τους ανάλυσης και επεξεργασίας. Εξαιρετικά αντιπροσωπευτικό παράδειγμα αυτού του είδους δεδομένων είναι οι σελίδες διαδικτύου, πάνω στις οποίες επιλέγουμε να προσαρμόσουμε την προτεινόμενη μεθοδολογία, και να κάνουμε την πειραματική της δοκιμή και μελέτη. Επιλέγουμε αυτού του είδους τα δεδομένα, καθώς είναι δυναμικά και άμεσα διαθέσιμα στους χρήστες αποδεσμευμένα από ενδεχόμενους περιορισμούς χρήσης.

Πιο συγκεκριμένα, με την προτεινόμενη μεθοδολογία, οι σελίδες διαδικτύου κατηγοριοποιούνται αυτοματοποιημένα και συνδυαστικά ως προς το θέμα τους και τον τύπο τους, όπως αυτός προκύπτει από τη δομή τους, διαδικασία που αποτελεί κρίσιμη παράμετρο για την αναζήτηση και ανάκτηση πληροφορίας στον παγκόσμιο ιστό. Παράλληλα, είναι γνωστό ότι τα δεδομένα διαδικτύου, μεταξύ αυτών και οι ιστοσελίδες, είναι δυναμικά. Δηλαδή αλλάζουν, ή μπορεί να αλλάζουν, ως προς το περιεχόμενό τους ή/και ως προς τη δομή τους, και μάλιστα με απρόβλεπτο τρόπο, σε απρόβλεπτο βαθμό και με απρόβλεπτο ρυθμό. Συνεπώς, ένα σύστημα αυτόματης κατηγοριοποίησης δεδομένων διαδικτύου θα πρέπει να λαμβάνει υπόψη και τη δυναμικότητα των σελίδων. Για τον λόγο αυτό, η προτεινόμενη μεθοδολογία αντιμετωπίζει τόσο την ετερογένεια των δεδομένων όσο και τον «άτακτο», ανομοιόμορφο ρυθμό και τρόπο αλλαγής μέσα στο χρόνο. Με άλλα λόγια, στο πλαίσιο της παρούσας διατριβής, σχεδιάζουμε, προτείνουμε και δοκιμάζουμε μία ολοκληρωμένη μεθοδολογία για την αυτοματοποιημένη και πολυδιάστατη, δηλαδή βάσει θεματικών στοιχείων και δομικών χαρακτηριστικών, κατηγοριοποίηση σελίδων διαδικτύου, η οποία παράλληλα λαμβάνει υπόψη το βαθμό και το ρυθμό αλλαγής των σελίδων, ώστε τα αποτελέσματα της κατηγοριοποίησης να παραμένουν επικαιροποιημένα.

4.2. Γενική Αρχιτεκτονική Προτεινόμενης Μεθοδολογίας

Στο πλαίσιο της παρούσας διατριβής, σχεδιάζεται και προτείνεται μία πρότυπη τεχνική πολυδιάστατης κατηγοριοποίησης σελίδων διαδικτύου. Συνοπτικά, η προτεινόμενη μεθοδολογία ολοκληρώνεται μέσα από τρεις ξεχωριστούς αλλά συμπληρωτικούς, αλγορίθμους, καθένας εκ των οποίων αποτελεί μία ολοκληρωμένη και αυτοτελή διαδικασία.

Πιο συγκεκριμένα, σκοπός του πρώτου αλγορίθμου (**ALGORITHM 1: Multi-Dimensional Page Classification**) είναι η πολυδιάστατη κατηγοριοποίηση των υπό επεξεργασία δεδομένων. Με τον όρο «*πολυδιάστατη*» (multi-dimensional), εννοούμε ότι λαμβάνονται υπόψη περισσότερες από μία παράμετροι, καθώς τα δεδομένα κατηγοριοποιούνται τόσο ως προς τη δομή, δομική κατηγοριοποίηση (**Procedure 1:** Structure-Based Classification) όσο και ως προς το περιεχόμενό τους, θεματική κατηγοριοποίηση (**Procedure 2:** Content-Based Classification). Σκοπός του δεύτερου αλγορίθμου (**ALGORITHM 2: Re-Classification based on Change Detection**), είναι ο έλεγχος της δυναμικότητας των δεδομένων, με απώτερο στόχο τον εντοπισμό εκείνων που χρειάζεται να επανακατηγοριοποιηθούν, και ως προς τι (δομή ή/και περιεχόμενο). Ολοκληρώνοντας την προτεινόμενη μεθοδολογία, σκοπός του τρίτου αλγορίθμου (**ALGORITHM 3: Optimized Re-Classification based on Change's Frequency Detection**) είναι η αποδοτικότερη λειτουργία του δεύτερου, υπολογίζοντας το ρυθμό αλλαγής των δεδομένων, και κατά συνέπεια τη συχνότητα επανεξέτασής τους ως προς το βαθμό αλλαγής και την ανάγκη για επανα-κατηγοριοποίησή τους.

Αναλυτικότερα, ο πρώτος αλγόριθμος (**ALGORITHM 1: Multi-Dimensional Page Classification**) ολοκληρώνεται μέσα από δύο επιμέρους *Διαδικασίες*, καθεμιά εκ των οποίων αποτελείται από δύο *Φάσεις*. Μέσα από την πρώτη διαδικασία (**Procedure 1:** Structure-based Classification) επιχειρείται η δομική κατηγοριοποίηση των δεδομένων, και κατά την πρώτη φάση της (**Phase 1:** *Page Type Recognition*) γίνεται ο διαχωρισμός των δεδομένων βάσει του τύπου τους, όπως αυτός προκύπτει από τη δομή τους, ενώ κατά τη δεύτερη φάση της (**Phase 2:** *Layered Page Classification*), με δεδομένο πια τον

τύπο τους και το δομικό τους χαρακτήρα, στόχος είναι η βαθύτερη κατηγοριοποίησή τους με βάση δομικά τους χαρακτηριστικά. Μέσα από τη δεύτερη διαδικασία (**Procedure 2:** Content-Based Classification) του ίδιου αλγορίθμου, επιχειρείται η θεματική κατηγοριοποίηση των υπό επεξεργασία δεδομένων, και αποτελείται επίσης από δύο διαφορετικές, αλλά συμπληρωματικές, φάσεις. Κατά την πρώτη φάση της (**Phase 1:** *Textual Elements Extraction*), εξάγονται τα θεματικά/κειμενικά στοιχεία εκείνα που χαρακτηρίζουν τα δεδομένα, ώστε να είναι δυνατή κατά τη δεύτερη φάση (**Phase 2:** *Theme Detection*) η/ο επιλογή/ανάθεση/εντοπισμός του θέματος για κάθε μία από τις σελίδες υπό επεξεργασία. Στη συνέχεια, μέσα από τον Αλγόριθμο 2 (**ALGORITHM 2: Re-Classification based on Change Detection**), πραγματοποιείται ο έλεγχος της δυναμικότητας των διαδικτυακών σελίδων. Με άλλα λόγια, σκοπός του Αλγορίθμου 2 είναι ο διαχωρισμός των σελίδων σε αυτές που παραμένουν ίδιες και σε εκείνες που παρουσιάζουν αλλαγές με την πάροδο του χρόνου, ο εντοπισμός των σημείων εκείνων όπου παρουσιάζονται οι αλλαγές, ο υπολογισμός τού βαθμού μεταβολής των σελίδων, και η λήψη απόφασης για την επανα-κατηγοριοποίησή τους ή όχι και με ποιον τρόπο (δομικά ή/και θεματικά). Η προτεινόμενη μεθοδολογία ολοκληρώνεται με τον Αλγόριθμο 3 (**ALGORITHM 3: Optimized Re-Classification based on Change's Frequency Detection**), ο οποίος εξετάζει το ρυθμό αλλαγής των δεδομένων, δηλαδή πόσο συχνά μεταβάλλονται ως προς το περιεχόμενο ή/και ως προς τη δομή τους. Μια τέτοια διαδικασία μπορεί να βελτιώσει τη λειτουργία του Αλγορίθμου 2, καθώς, ελέγχοντας το ρυθμό μεταβολής των δεδομένων, μπορεί να σχεδιαστεί καλύτερα η πολιτική επανα-κατηγοριοποίησης με σκοπό να εξοικονομούνται χρόνος και υπολογιστικό κόστος.

Συνοψίζοντας, η προτεινόμενη μεθοδολογία στοχεύει στην πολυδιάστατη κατηγοριοποίηση των σελίδων διαδικτύου και, δεδομένης της δυναμικότητας που τις χαρακτηρίζει, εντοπίζει ποιες από αυτές χρειάζονται επανα-κατηγοριοποίηση και ως προς τι, αλλά και πόσο συχνά χρειάζεται να επανεξεταστούν. Στη συνέχεια του Κεφαλαίου αυτού, δίνονται οι ψευδοκώδικες των αλγορίθμων που αποτυπώνουν την προτεινόμενη μεθοδολογία, η αναλυτική περιγραφή των βημάτων που περιλαμβάνει κάθε διαδικασία, καθώς και η επεξήγηση των λεκτικών/συμβόλων που αντικαθιστούν φράσεις μέσα στους ψευδοκώδικες. Συνολικά, οι ψευδοκώδικες των

αλγορίθμων που περιλαμβάνει η προτεινόμενη μεθοδολογία ως ένα ενιαίο σύνολο, παρουσιάζονται στο **Παράρτημα 1:** Ψευδοκώδικες αλγορίθμων προτεινόμενης μεθοδολογίας. Ομοίως και η επεξήγηση των λεκτικών που χρησιμοποιούνται σε αυτούς, **Παράρτημα 2:** Ορολογία και λεκτικά αλγορίθμων (ελληνικά), όπως και η σχηματική απεικόνιση της μεθοδολογίας, **Παράρτημα 3:** Σχηματική απεικόνιση προτεινόμενης μεθοδολογίας.

4.3. Αναλυτική Περιγραφή Μεθοδολογίας

4.3.1. Πολυδιάστατη Κατηγοριοποίηση Σελίδων Διαδικτύου

Σκοπός του πρώτου αλγορίθμου της προτεινόμενης μεθοδολογίας (**ALGORITHM 1: Multi-Dimensional Page Classification**) είναι η πολυδιάστατη κατηγοριοποίηση των δεδομένων υπό επεξεργασία. Στο πλαίσιο της παρούσας διατριβής, όπως έχει ήδη αναφερθεί, τα δεδομένα αυτά είναι τα έγγραφα του παγκόσμιου ιστού, δηλαδή οι διαδικτυακές σελίδες. Πιο συγκεκριμένα, μέσα από τις επιμέρους διαδικασίες του Αλγορίθμου 1, επιδιώκεται η κατηγοριοποίηση των δεδομένων με βάση τόσο τα δομικά όσο και τα κειμενικά τους χαρακτηριστικά.

Ξεκινώντας από την περιγραφή του αλγορίθμου δομικής κατηγοριοποίησης, για τη σχεδίασή του υιοθετούμε οπτική από την πλευρά της ανάκτησης πληροφορίας, και βασιζόμαστε στην εξής **βασική θεώρηση**: *με δεδομένες τις κατηγορίες των πληροφοριακών αιτημάτων των χρηστών του διαδικτύου [39], λογικά υπάρχουν και οι αντίστοιχες κατηγορίες σελίδων που τα ικανοποιούν. Παράλληλα, παίρνουμε σαν δεδομένο το γεγονός ότι οι σελίδες με παρόμοια δομή, είναι του ίδιου τύπου, δηλαδή ανήκουν στην ίδια κατηγορία. Κατά συνέπεια, οδηγούμαστε στο συμπέρασμα ότι οι σελίδες που ικανοποιούν το εκάστοτε αίτημα των χρηστών, έχουν κοινά χαρακτηριστικά ως προς τη δομή τους.*

Συνοπτικά, ο Αλγόριθμος 1 ολοκληρώνεται μέσω δύο επιμέρους διαδικασιών, καθεμιά εκ των οποίων αποτελείται από δύο φάσεις. Η πρώτη διαδικασία (**Procedure 1:** Structure-Based Classification) έχει σκοπό τη δομική κατηγοριοποίηση των δεδομένων, και κατά την πρώτη φάση (**Phase 1:** *Page Type Recognition*) γίνεται ο

διαχωρισμός των δεδομένων βάσει του τύπου τους, ενώ κατά τη δεύτερη (**Phase 2:** *Layered Page Classification*), με δεδομένο τον τύπο τους και το δομικό τους χαρακτήρα, επιχειρείται βαθύτερη κατηγοριοποίηση. Η δεύτερη διαδικασία (**Procedure 2:** Content-based Classification) έχει στόχο τη θεματική κατηγοριοποίηση των δεδομένων, και επίσης ολοκληρώνεται μέσα από δύο διαφορετικές φάσεις. Κατά την πρώτη φάση (**Phase1:** *Textual elements extraction*), εξάγονται τα θεματικά/κειμενικά στοιχεία που χαρακτηρίζουν τα δεδομένα, και έχουμε αξιολογήσει ως αντιπροσωπευτικά, ώστε να είναι δυνατή κατά τη δεύτερη φάση (**Phase2:** *Theme Selection*) η επιλογή/ανάθεση θέματος για κάθε ένα από τα δεδομένα υπό επεξεργασία.

Στην εικόνα που ακολουθεί, Εικόνα 3, φαίνεται ο ψευδοκώδικας της πρώτης φάσης της πρώτης διαδικασίας του αλγορίθμου κατηγοριοποίησης, κατά την οποία οι σελίδες διαδικτύου χαρακτηρίζονται με βάση τον τύπο τους, όπως αυτός προκύπτει από τη δομή των ίδιων των σελίδων.

ALGORITHM 1: MULTI-DIMENSIONAL PAGE CLASSIFICATION

```
1   PROCEDURE 1: STRUCTURE-BASED CLASSIFICATION
2     PHASE 1: PAGE TYPE RECOGNITION
3     Input: P, tokenizer, T(trans), Text-to-Link-Analyzer, (t)
4         for every P
5           look for t(trans) appearing as link
6             if any
7                 tag P as P(transactional)
8             Else
9                 compute word tokens to links ratio (R)
10                if R ≥ t
11                    tag P as P(informational)
12                Else
13                    tag P as P(navigational)
14                    End
15        End
16    Output: P(transactional), P(navigational), P(informational)
```

Εικόνα 3: ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 1: Structure-Based Classification, **Phase 1:** *Page Type Recognition*

Συνοπτικά, κατά την πρώτη φάση της πρώτης διαδικασίας, (**Phase1:** *Page Type Recognition*), ο αλγόριθμος, με βάση τα δομικά χαρακτηριστικά της κάθε σελίδας, τις διαχωρίζει με βάση τον τύπο τους, ο οποίος προκύπτει από το σκοπό που εξυπηρετούν σύμφωνα με τη θεώρηση που αναφέρουμε ήδη πιο πάνω. Τα δεδομένα που δίνονται ως *δεδομένα εισόδου* (*Input*) είναι οι σελίδες προς κατηγοριοποίηση (*(P)*), ένα εργαλείο επεξεργασίας των λεκτικών μονάδων που απαρτίζουν ένα κείμενο

(*tokenizer*), μία λίστα με τους όρους διάδρασης (*T(trans)*), ένα εργαλείο για τον υπολογισμό της αναλογίας κειμένου - (υπερ-)συνδέσμων (*Text-to-Link-Analyzer*), με άλλα λόγια τον λόγο του ποσοστού της κειμενικής πληροφορίας που περιλαμβάνει μια σελίδα προς το ποσοστό των συνδέσμων που περιλαμβάνει η ίδια σελίδα, και ένα όριο (*t*) στο οποίο βασίζεται ο Αλγόριθμος για το διαχωρισμό των σελίδων σε πλοήγησης και πληροφοριακές με δεδομένη την προαναφερθείσα αναλογία.

Αναλυτικά, κατά το πρώτο βήμα της διαδικασίας αυτής, ο αλγόριθμος εξετάζει αν στο κυρίως σώμα της σελίδας υπάρχει τουλάχιστον ένας όρος διάδρασης (*transactional term (t(trans))*), ο οποίος εμφανίζεται με τη μορφή συνδέσμου. Όροι διάδρασης είναι οι όροι εκείνοι, συχνότερα ρήματα, που βρίσκονται στο σώμα της σελίδας και δηλώνουν κάποια ενέργεια μέσα από την οποία ο χρήστης μπορεί να απολαύσει την παροχή κάποιας υπηρεσίας ή να αποκτήσει κάποιο προϊόν, δωρεάν ή όχι [39]. Παράλληλα, για τον αλγόριθμό μας, είναι απαραίτητο ο όρος αυτός να έχει τη μορφή συνδέσμου, καθώς αυτό είναι το χαρακτηριστικό εκείνο που υποδεικνύει πως, επιλέγοντάς τον, ο χρήστης οδηγείται στην εκτέλεση της ενέργειας διάδρασης/συναλλαγής που δηλώνει ο αντίστοιχος όρος. Αν υπάρχει έστω και ένας όρος διάδρασης από όσους έχουν δοθεί υπό μορφή λίστας στα δεδομένα εισόδου, τότε ο αλγόριθμος χαρακτηρίζει τη σελίδα αυτή *σελίδα διάδρασης* (*transactional page*).

Ο λόγος για τον οποίο επιλέγεται αυτό το βήμα να είναι το πρώτο αυτής της διαδικασίας, είναι γιατί η ύπαρξη ενός ή περισσότερων όρων διάδρασης, (*t(trans)*), είναι ικανή και αρκετή για να τεκμηριώσει το χαρακτηρισμό μιας σελίδας ως σελίδας διάδρασης. Με αυτόν τον τρόπο, καθιστούμε τον αλγόριθμό μας ικανό, εύκολα και γρήγορα να διαχωρίσει τις σελίδες εκείνες που ανήκουν στη μία εκ των τριών κατηγοριών σελίδων βάσει δομής. Οι υπόλοιπες σελίδες στέλνονται στο επόμενο βήμα, όπου γίνεται ο διαχωρισμός τους μεταξύ *σελίδων πλοήγησης* (*navigational pages*) και *πληροφοριακών* (*informational pages*). Για να είναι δυνατός αυτός ο διαχωρισμός, ο αλγόριθμός μας βασίζεται στην αναλογία κειμένου - (υπερ-)συνδέσμων που παρουσιάζει η εκάστοτε σελίδα. Κι αυτό γιατί οι πληροφοριακές σελίδες αποτελούνται κυρίως από κείμενο, ενώ οι σελίδες πλοήγησης περιέχουν κυρίως συνδέσμους ή/και (υπερ-)συνδέσμους [92]. Έτσι, κατά το δεύτερο βήμα του

αλγορίθμου κατηγοριοποίησης, υπολογίζεται σε τι ποσοστό η κάθε μία από τις σελίδες που εξετάζονται περιλαμβάνουν κείμενο και σε τι ποσοστό (υπερ-)συνδέσμους. Σε αυτό το σημείο, ο αλγόριθμος αξιοποιεί το όριο που έχει δοθεί στα δεδομένα εισόδου (*t*), και αν η αναλογία κειμένου - (υπερ-)συνδέσμων είναι μεγαλύτερη του ορίου αυτού, τότε η σελίδα χαρακτηρίζεται πληροφοριακή. Αν η αναλογία κειμένου - (υπερ-)συνδέσμων είναι μικρότερη του ορίου, τότε η ιστοσελίδα χαρακτηρίζεται πλοήγησης. Αποτέλεσμα αυτής της διαδικασίας είναι ο χαρακτηρισμός κάθε σελίδας με βάση τον τύπο της, όπως αυτός προκύπτει από τη δομή της.

Κατά τη δεύτερη φάση της ίδιας διαδικασίας (**ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 1:** Structure-Based Classification, **Phase 2:** *Layered Page Classification*), όπως αναφέρεται και πιο πάνω, γίνεται βαθύτερη δομική κατηγοριοποίηση των σελίδων, με δεδομένο πλέον τον τύπο τους. Η κατηγοριοποίηση αυτή εκτελείται μόνο για τους δύο εκ των τριών τύπων σελίδων, δηλαδή για τις σελίδες πλοήγησης και διάδρασης, καθώς οι πληροφοριακές δεν παρουσιάζουν δομή τέτοια από την οποία θα μπορούσε να απορρέει κάποια βαθύτερη δομική κατηγοριοποίηση. Επίσης, διευκρινίζεται ότι, η δομική κατηγοριοποίηση των σελίδων πλοήγησης, και αυτή των σελίδων διάδρασης, είναι διεργασίες οι οποίες μπορούν να εκτελεστούν παράλληλα, μιας και πρόκειται για διεργασίες ανεξάρτητες και αυτοτελείς. Στην εικόνα που ακολουθεί, Εικόνα 4, φαίνεται ο ψευδοκώδικας της δεύτερης φάσης της πρώτης διαδικασίας τού αλγορίθμου κατηγοριοποίησης.

```
17  PHASE 2: LAYERED PAGE CLASSIFICATION GIVEN THE TYPE
18  Input: P(transactional), P(navigational), T(corr), T(payment) D(top), LinkC, (h)
19      for every P(transactional)
20          map P(transactional) to the Table(corr)
21              for every mapping found
22                  count occurrences and tag P(transactional) with the category of max occurrence
23              Else
24                  look for t(payment) appearing as link
25                      if t(payment) ≥ 1
26                          tag P(transactional) as "not-free"
27                      Else
28                          tag P(transactional) as "free"
29                          End
30      for every P(navigational) starting after "http(s)://"
31          count the number of "/" in url
32              if "/" ≥ h
```

```
33                      tag P(navigational) as "WebPage" and
34                      set the number of "/" as depth value
35                      End
36                  else
37                      tag P(navigational) as "HomePage" and
38                      map the HomePage suffix to the D(top)
39                        if there is a mapping
40                            tag HomePage with the suffix meaning
41                            End
42                        Else
43                            validate url against LinkC
44                              for every valid link
45                                if internal
46                                  set the number of (/) as depth value
47                                  End
48                                Else
49                                  send P(navigational) to Procedure1
50                                  End
51          End
52  Output: P(transactional), P(navigational), P(informational)
```

Εικόνα 4: ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 1: Structure-Based Classification, **Phase 2:** *Layered Page Classification*

Τα δεδομένα εισόδου σε αυτή τη φάση του αλγορίθμου είναι οι σελίδες διάδρασης (*P(transactional)*), οι σελίδες πλοήγησης (*P(navigational)*), ένας πίνακας όπου υπάρχει αντιστοίχιση όρου διάδρασης και κατηγορίας στην οποία ανήκει η ενέργεια (*T(corr)*), ένας πίνακας με τους όρους που (υπο-)δηλώνουν απαιτούμενη οικονομική συναλλαγή για την ολοκλήρωση της ενέργειας διάδρασης (*T(payment)*), ένας πίνακας όπου φαίνεται η αντιστοιχία των καταλήξεων που έχουν οι ενιαίοι εντοπιστές πόρων (*URLs*) των σελίδων, (*D(top)*), ένα εργαλείο για την καταμέτρηση των εσωτερικών και εξωτερικών συνδέσμων που περιέχονται σε μια σελίδα, (*LinkC*), και τέλος ένα όριο (*h*) για τον εντοπισμό των σελίδων που αποτελούν την αρχική ενός ιστοτόπου (*homepages*). Κατά τη δεύτερη φάση του πρώτου αλγορίθμου (**Phase2:** *Layered Page Classification*), η πρώτη κατηγοριοποίηση που παρουσιάζεται είναι αυτή των σελίδων διάδρασης. Αυτό, όπως προκύπτει από τα παραπάνω, δεν δηλώνει προτεραιότητα εκτέλεσης της εν λόγω διαδικασίας. Οι επιμέρους δομικές κατηγοριοποιήσεις των σελίδων διάδρασης και πλοήγησης είναι διαδικασίες ανεξάρτητες και αυτοτελείς και εκτελούνται παράλληλα.

Για την βαθύτερη δομική κατηγοριοποίηση των ιστοσελίδων διάδρασης, κατά το πρώτο βήμα, και στο πλαίσιο του αλγορίθμου, αντιστοιχίζονται οι όροι διάδρασης στον πίνακα, όπου φαίνονται οι κατηγορίες στις οποίες μπορεί να ανήκει η κάθε

ενέργεια. Εντοπίζοντας την κατηγορία ενέργειας στην οποία ανήκουν οι περισσότεροι - από τους περιεχόμενους στην υπό εξέταση σελίδα - όροι διάδρασης, την ορίζουμε ως κατηγορία της σελίδας. Η διεργασία αυτή, είναι σημαντική, αφού ο όρος διάδρασης αποτελεί ένα πολύ ισχυρό χαρακτηριστικό της σελίδας στην οποία συναντάται, και μας δίνει τη δυνατότητα να δώσουμε και έναν επιπλέον, πιο ειδικό χαρακτηρισμό στη σελίδα.

Στη συνέχεια, και για τις ίδιες σελίδες, ο αλγόριθμος ελέγχει αν υπάρχει κάποιος όρος που να (υπό-)δηλώνει την απαίτηση οικονομικής συναλλαγής, όπως αυτοί εμφανίζονται στον αντίστοιχο πίνακα (*T(payment)*) που έχει δοθεί ως δεδομένο εισόδου. Αν υπάρχει έστω και ένας τέτοιος όρος, τότε η συναλλαγή που μπορούν να κάνουν οι χρήστες μέσω της υπό επεξεργασία σελίδας χαρακτηρίζεται «*NotFree*», διαφορετικά χαρακτηρίζεται «*Free*», δηλαδή δωρεάν. Ο χαρακτηρισμός αυτός, είναι άξιος αναφοράς και μελέτης στο πλαίσιο της μεθοδολογίας μας, αφού όχι μόνο υποδεικνύει το οικονομικό πλαίσιο της συναλλαγής, αλλά αποτελεί και απόδειξη αυτής. Με αυτό το διαχωρισμό, ολοκληρώνεται η βαθύτερη βάσει δομής κατηγοριοποίηση των σελίδων διάδρασης.

Στο πλαίσιο άλλης διαδικασίας που εκτελείται παράλληλα, κατηγοριοποιούνται βαθύτερα και βάσει δομικών χαρακτηριστικών οι σελίδες πλοήγησης. Κατά το πρώτο βήμα, ο αλγόριθμος καλείται να μετρήσει το πλήθος των *slashes* (/) που εμφανίζονται στον ενιαίο εντοπιστή πόρων (*URL*) της κάθε σελίδας, ξεκινώντας το μέτρημα μετά το πρόθεμα «*http(s)://*». Ο λόγος για τον οποίο επιλέγουμε να λάβουμε υπόψη μας αυτό το στοιχείο, είναι γιατί βάσει αυτού ο αλγόριθμός μας θα μπορέσει να αποφασίσει αν πρόκειται για την αρχική σελίδα ενός ιστοτόπου ή για κάποια άλλη που βρίσκεται πιο βαθιά στο εσωτερικό της δομής του, δηλαδή πρόκειται για ιστοσελίδα. Μελετώντας τους κανόνες σύνταξης μιας διεύθυνσης σελίδας τού διαδικτύου [13], φαίνεται πως το δεδομένο αυτό δείχνει σε ποιο βάθος ενός ιστοτόπου βρίσκεται μια ιστοσελίδα. Παράλληλα, ο λόγος που επιλέγουμε να μην τα μετρήσουμε όλα τα *slashes*, αλλά να ξεκινήσουμε να μετράμε μετά το πρόθεμα, είναι γιατί τα *slashes* που περιέχονται στο πρόθεμα ενός *URL* δεν παίζουν ρόλο όσον αφορά το βάθος στο οποίο συναντάμε μια σελίδα. Αντιθέτως, αν το συμπεριλαμβάναμε στον υπολογισμό, θα ήταν παραπλανητικό για την απόφαση που θα πάρει ο αλγόριθμος, καθώς το

πρόθεμα αυτό υπάρχει στις περισσότερες διευθύνσεις σελίδων διαδικτύου. Εναλλακτικά, θα έπρεπε να αφαιρέσουμε δύο από το πλήθος που θα εντόπιζε ο αλγόριθμος, γεγονός που επίσης θα μπορούσε να μας οδηγήσει σε λάθος συμπέρασμα, μιας και δεν έχουν όλα τα *URLs* το συγκεκριμένο πρόθεμα απαραιτήτως, και θα χρειαζόταν η διεξαγωγή ενός ακόμη βήματος μέσα στον αλγόριθμο, γεγονός που θα απαιτούσε περισσότερο χρόνο και επιπλέον υπολογιστικούς πόρους. Σύμφωνα με τη βιβλιογραφία, η δομή και τα γνωρίσματα των URLs, μπορούν να συντελέσουν στην κατηγοριοποίηση σελίδων διαδικτύου [28] [6].

Έχοντας μετρήσει ο αλγόριθμος τα *slashes*, και έχοντας σαν σημείο αναφοράς το όριο που έχει λάβει ως δεδομένο εισόδου, αποφασίζει αν η σελίδα που εξετάζει αποτελεί την αρχική ενός ιστοτόπου ή κάποια που βρίσκεται βαθύτερα στο εσωτερικό του. Πιο συγκεκριμένα, αν το πλήθος των *slashes* είναι μεγαλύτερο του ορίου, τότε πρόκειται για κάποια ιστοσελίδα στο εσωτερικό ενός ιστοτόπου (*webpage*), διαφορετικά πρόκειται για την αρχική σελίδα (*homepage*). Στην πρώτη περίπτωση, ο αλγόριθμος καταχωρεί το πλήθος των *slashes* ως τιμή βάθους της ιστοσελίδας και ολοκληρώνεται εκεί η δομική της κατηγοριοποίηση. Στην περίπτωση μιας αρχικής σελίδας πλοήγησης, ο αλγόριθμος εντοπίζει την κατάληξη του ενιαίου εντοπιστή πόρου (*URL*) και καταχωρεί την κατηγορία στην οποία ανήκει, αφού κάνει την αντιστοιχία στον σχετικό πίνακα (*D(top)*).

Συμπληρωματικά, προκειμένου να είναι ολοκληρωμένη η επεξεργασία των δεδομένων, ο αλγόριθμος εξετάζει τους περιεχόμενους - σε κάθε σελίδα που χαρακτηρίζεται αρχική - συνδέσμους, αξιοποιώντας ένα σχετικό εργαλείο (*LinkC*). Για κάθε έγκυρο σύνδεσμο, ελέγχει αν αυτός είναι εσωτερικός, δηλαδή οδηγεί σε κάποια άλλη ιστοσελίδα του ίδιου ιστοτόπου, ή αν πρόκειται για υπερ-σύνδεσμο, δηλαδή οδηγεί σε κάποια σελίδα εξωτερική, είτε είναι αρχική είτε όχι, που ανήκει σε κάποιον άλλο ιστότοπο. Στην πρώτη περίπτωση, υπολογίζονται τα *slashes*, όπως πιο πάνω, και καταχωρούνται ως τιμή βάθους στο οποίο βρίσκεται η ιστοσελίδα στην οποία οδηγεί ο σύνδεσμος. Στη δεύτερη περίπτωση, όταν πρόκειται για κάποιον εξωτερικό σύνδεσμο, δηλαδή υπερ-σύνδεσμο, αυτός στέλνεται στην αρχή του πρώτου αλγορίθμου για μια ολοκληρωμένη κατηγοριοποίηση. Αποτέλεσμα όλων των

παραπάνω διαδικασιών είναι η κατηγοριοποίηση των εγγράφων του διαδικτύου με βάση τα δομικά τους χαρακτηριστικά. Με άλλα λόγια, μέσα από αυτές, ο αλγόριθμος καταφέρνει να χαρακτηρίσει κάθε σελίδα διαδικτύου ως προς τον τύπο της.

Στη συνέχεια, προκειμένου να ολοκληρωθεί η πολυδιάστατη κατηγοριοποίηση των ιστοσελίδων, ο Αλγόριθμος 1 εκτελεί τη θεματική κατηγοριοποίηση των τελευταίων, βάσει των κειμενικών τους χαρακτηριστικών (**ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 2**: Content-Based Classification).

Η θεματική κατηγοριοποίηση των σελίδων ολοκληρώνεται σε δύο φάσεις, οι οποίες είναι άρρηκτα συνδεδεμένες μεταξύ τους και εκτελούνται η μία μετά την άλλη, όπως παρουσιάζονται παρακάτω. Κατά την πρώτη φάση (**Phase 1:** *Textual Element Extraction*), γίνεται η εξαγωγή των κειμενικών χαρακτηριστικών της κάθε σελίδας, ενώ κατά τη δεύτερη φάση (**Phase2:** *Theme Detection*) γίνεται η τελική επιλογή του θέματος. Στην εικόνα που ακολουθεί, Εικόνα 5, φαίνεται ο ψευδοκώδικας της πρώτης φάσης τής δεύτερης διαδικασίας του αλγορίθμου κατηγοριοποίησης.

ALGORITHM 1: CONTENT-BASED CLASSIFICATION

```
1   PROCEDURE 2: CONTENT-BASED CLASSIFICATION
2     PHASE 1: TEXTUAL ELEMENTS EXTRACTION
3     Input: P
4       for each P
5         search for anchor title in url
6           if any
7             tag as "P's anchorTitle"
8             End
9           Else
10            search for title in text body
11              if any
12                tag as "P's textTitle"
13                End
14      End
15    Output: P tagged with Textual elements
```

Εικόνα 5: ALGORITHM 1: Multi-Dimensional Page Classification, **Procedure 2:** Content-Based Classification, **Phase 1:** *Textual Elements Extraction*

Πιο αναλυτικά, κατά την πρώτη φάση της δεύτερης διαδικασίας (**Phase 1:** *Textual Element extraction*), όπου σκοπός είναι η εξαγωγή των κειμενικών χαρακτηριστικών των σελίδων, ως δεδομένα εισόδου δίνονται οι σελίδες προς κατηγοριοποίηση. Κατά το πρώτο βήμα, ο αλγόριθμός μας ελέγχει και αποθηκεύει, όπου υπάρχει, το *Anchor Title* για κάθε σελίδα. Όταν αναφερόμαστε στο *P's anchorTitle* εννοούμε τις λέξεις

εκείνες που εμφανίζονται ως σύνδεσμος και μας οδηγούν στη σελίδα. Αλλιώς, από τη σκοπιά του χρήστη, είναι ο τίτλος που φαίνεται κατά την αποθήκευση της σελίδας στους σελιδοδείκτες του φυλλομετρητή. Πιο απλά, πρόκειται για τις λέξεις που εμφανίζονται ως τίτλος περιγραφής της σελίδας, και στον κώδικα html της σελίδας ορίζεται μέσα από το γνώρισμα *<a href>*, από όπου μπορεί και να αντληθεί. Το κείμενο αυτό, όπου το συναντάει ο αλγόριθμος, το αποθηκεύει με τη συγκεκριμένη ετικέτα.

Το συγκεκριμένο στοιχείο αποτελεί πολύ ισχυρή ένδειξη θέματος, αφού συχνά ταυτίζεται με το html element *<pageTitle>* ή *<contentTitle>*, και όταν δεν ταυτίζεται είναι όμοιο κατά τουλάχιστον 50%. Πολύ σημαντικό είναι, επίσης, το γεγονός ότι το συναντάμε σε όλες τις σελίδες, άρα πρόκειται για ένα κοινό στοιχείο. Επίσης, πρόκειται για στοιχείο που αξιοποιείται για την κατηγοριοποίηση των ιστοσελίδων, αλλά και ευρύτερα για την κατηγοριοποίηση δεδομένων διαδικτύου [30][65][26][48].

Στη συνέχεια, ο αλγόριθμος εντοπίζει τον τίτλο, *Title*, που μπορεί να έχει η κάθε σελίδα στο κυρίως σώμα της. Πρόκειται για μία ακόμα ισχυρή ένδειξη για το θέμα μιας σελίδας. Επομένως, όπου το συναντάει ο Αλγόριθμός μας, το αποθηκεύει δίνοντας την ετικέτα «title». Το στοιχείο «*P's textTitle*», από την άλλη, είναι ο τίτλος που εμφανίζεται (όπου υπάρχει) στο κυρίως σώμα της σελίδας και στον hmtl κώδικα της σελίδας ορίζεται με τις ετικέτες *<h1>* έως *<h6>*, ανάλογα με την σπουδαιότητα του τίτλου [VIII]. Δηλαδή, η ετικέτα *<h1>* αφορά τον κύριο τίτλο, η ετικέτα *<h2>* τον υπότιτλο κ.ο.κ. [IX]. Στο σημείο αυτό, ολοκληρώνεται η πρώτη φάση του αλγορίθμου θεματικής κατηγοριοποίσης, και το αποτέλεσμα που έχουμε ύστερα από την ολοκλήρωσή της, είναι η ετικετοποίηση κάθε σελίδας με τουλάχιστον ένα ισχυρό χαρακτηριστικό θέματος, αν όχι και με τα δύο προαναφερθέντα.

Ύστερα, ακολουθεί η δεύτερη Φάση της θεματικής κατηγοριοποίησης, όπου γίνεται η τελική επιλογή του θέματος για κάθε σελίδα (**ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 2:** Content-Based Classification, **Phase 2:** *Theme Detection*). Σε αυτή τη φάση, τα δεδομένα *εισόδου* είναι τα θεματικά χαρακτηριστικά στοιχεία που έχουν εξαχθεί κατά την πρώτη φάση (*(P's anchorTitle), (P's textTitle)*), οι κατηγορίες της ηλεκτρονικής εγκυκλοπαίδειας Wikipedia με τα περιεχόμενά τους, μια μετρική

ομοιότητας (*n-gram similarity metric*), ένα εργαλείο για την υπολογιστική επισήμανση των μερών του λόγου ενός κειμένου γραμμένου σε φυσική γλώσσα (*PoS-Tagger*), έναν συντακτικό αναλυτή φυσικής γλώσσας (*Parser*), το δίκτυο σημασιολογικά συσχετισμένων όρων WordNet, έναν λημματοποιητή (*lemmatizer*), η μετρική (*TF*IDF*), και το όριο (*n*) για την εξαγωγή λέξεων-κλειδιών.

Στην εικόνα που ακολουθεί, Εικόνα 6, φαίνεται ο ψευδοκώδικας της δεύτερης φάσης τής δεύτερης διαδικασίας του πρώτου αλγορίθμου.

```
16  PHASE 2: THEME DETECTION
17  Input: (P's anchorTitle), (P's textTitle), WebPage Word Counter, PoS-Tagger, Parser, WordNet, lemmatizer,
    (TF*IDF), (n), WP contents, WP articles
18      for each P look for common terms between P's anchorTitle and P's textTitle
19          if found
20              use common terms as the thematic term(-s) to tag P and map thematic(s) term(s) to WP contents
21                  for every mapping found
22                      count occurrences and tag P with the category of max occurrence
23                      End
24                  Else
25                      search for WP article titled with the wider thematic term
26                      map article's categories to WP contents begging from the first one
27                      stop when a mapping is found and tag P with the category
28                      End
29          Else
30              PoS-tag and lemmatize P's text and extract the first n-appearing keywords
31              check for overlapping terms between P's keywords and (P's anchor title and P's text title)
32                  if found
33                      use overlapping terms as the thematic term(-s) to tag P and map thematic(s) term(s) to WP contents
34                          for every mapping found
35                              count occurrences and tag P with the category of max occurrence
36                              End
37                          Else
38                              search for WP article titled with the wider thematic term
39                              map article's categories to WP contents begging from the first one
40                              stop when a mapping is found and tag P with the category
41                              End
42          Else
43              map P's first n-appearing keywords to WordNet and look for common senses between P's keywords and
                (P's anchor title and P's text title)
44                  if found
45                      use terms of common senses as the thematic term(-s) to tag P and map thematic(s) term(s) to WP
                        contents
46                          for every mapping found
47                              count occurrences and tag P with the category of max occurrence
48                              End
49                          Else
50                              search for WP article titled with the wider thematic term
51                              map article's categories to WP contents begging from the first one
52                              stop when a mapping is found and tag P with the category
53                              End
54              Else
55                  tag P as unknown category (Punknown)
56                  End
57      end
58  Output: Thematically classified P
59  Output: Multi-Dimensionally classified P
```

Εικόνα 6: ALGORITHM 1: Multi-Dimensional Page Classification, **Procedure 2:** Content-Based Classification, **Phase 2:** *Theme Detection*

Για κάθε σελίδα προς επεξεργασία, ο αλγόριθμός μας αρχικώς ελέγχει αν υπάρχουν κοινοί όροι μεταξύ των στοιχείων «*P's anchorTitle*» και «*P's textTitle*». Αν υπάρχουν, τότε αυτοί οι κοινοί όροι θεωρούνται θεματικοί όροι της σελίδας. Ο λόγος για τον οποίο ο αλγόριθμός μας θεωρεί τους κοινούς αυτούς όρους αρκετούς και ικανούς για να μας οδηγήσουν στο θέμα της σελίδας, είναι το γεγονός ότι, όπως αναφέρεται και πιο πάνω, έτσι κι αλλιώς τα δύο αυτά στοιχεία, ακόμα και μεμονωμένα, αποτελούν πολύ ισχυρή ένδειξη θέματος, πόσο μάλλον όταν αυτά εμφανίζουν κοινούς όρους.

Στην περίπτωση που δεν υπάρχουν κοινές λέξεις μεταξύ των δύο αυτών στοιχείων, ο αλγόριθμός μας, εκτελεί το επόμενο κατά σειρά βήμα, κατά το οποίο καλείται να αναλύσει μορφοσυντακτικά το περιεχόμενο και να εξάγει τις λέξεις-κλειδιά (*keywords*) από το σώμα της κάθε σελίδας υπό επεξεργασία. Στο σημείο αυτό έχουμε ορίσει και ένα όριο, ώστε να μην λαμβάνονται υπόψη όλες οι λέξεις-κλειδιά, αλλά οι (*n*) πρώτες με βάση τη σειρά εμφάνισης, που είναι και οι πιο ισχυρές. Ύστερα, ο αλγόριθμος ελέγχει αν υπάρχουν κοινοί όροι μεταξύ λέξεων-κλειδιών και «*P's anchorTitle*» ή/και «*P's textTitle*». Αν ικανοποιείται αυτή η συνθήκη, τότε ορίζονται οι κοινοί όροι ως θεματικοί όροι της σελίδας.

Αν όμως, ακόμα η σελίδα παραμένει χωρίς θεματικούς όρους, ο αλγόριθμός μας, αξιοποιώντας τη λεξικογραφική οντολογία τού WordNet, ελέγχει αν κάποιο/α από τις λέξεις-κλειδιά έχουν την ίδια σημασία με κάποιον/ους από τους όρους τού «*P's anchorTitle*» ή/και του «*P's textTitle*». Αν συμβαίνει αυτό, επιλέγονται αυτοί οι όροι ως θεματικοί όροι τις σελίδας, διαφορετικά η σελίδα παραμένει χωρίς θεματικούς όρους και άρα χωρίς θέμα και χαρακτηρίζεται (*Punknown*).

Στη συνέχεια, με βάση αυτούς τους θεματικούς όρους που έχουν προκύψει σε κάποιο από τα παραπάνω βήματα, και αξιοποιώντας τις κατηγορίες της Wikipedia, προκύπτει η θεματική κατηγορία κάθε σελίδας. Σύμφωνα με τα βήματα του αλγορίθμου, προκειμένου να ομαδοποιηθούν και να ονοματοδοτηθούν τα θέματα των σελίδων υπό επεξεργασία, αντιστοιχώνται οι κοινοί θεματικοί όροι που αναφέρονται πιο πάνω στα περιεχόμενα των κατηγοριών της Wikipedia, και η σελίδα υιοθετεί την

κατηγορία στην οποία ανήκουν οι περισσότεροι από τους θεματικούς της όρους. Αν αυτοί οι όροι δεν βρίσκονται στα περιεχόμενα των κατηγοριών, ο αλγόριθμός μας ορίζει πως επιλέγεται ο ευρύτερος από τους θεματικούς όρους της σελίδας και γίνεται αναζήτηση για σχετικό άρθρο στη Wikipedia με τίτλο αυτόν τον όρο. Έτσι, σε αυτή την περίπτωση, η σελίδα υιοθετεί την κατηγορία στην οποία ανήκει το άρθρο αυτό. Για τον εντοπισμό της κατηγορίας στην οποία ανήκει το άρθρο και άρα η σελίδα, στηρίζεται στις κατηγορίες που εμφανίζονται ως ετικέτες στο κάτω μέρος του άρθρου, και οι οποίες είναι τοποθετημένες από τους επιμελητές με σειρά σχετικότητας.

Οι λόγοι για τους οποίους επιλέγουμε να στηριχτούμε στα περιεχόμενα των κατηγοριών της Wikipedia για να ομαδοποιήσουμε και να ονοματοδοτήσουμε τα θέματα των σελίδων υπό επεξεργασία, είναι αφενός επειδή η Wikipedia παρουσιάζει κάποιες ομοιότητες με την μεθοδολογία μας και αφετέρου επειδή συγκεντρώνει ορισμένα σημαντικά για τη δουλειά μας χαρακτηριστικά. Συγκεκριμένα, η ηλεκτρονική εγκυκλοπαίδεια Wikipedia χαρακτηρίζεται από ιεραρχική δομή, ενώ και τα βήματα του αλγορίθμου μας έχουν επίσης ιεραρχική πορεία. Επίσης, πρόκειται για μια ευρέως αποδεκτή εγκυκλοπαίδεια και ως προς το περιεχόμενό της και ως προς τη δομή της ως δομή οντολογίας. Μάλιστα, η οργάνωση των κατηγοριών προέκυψε από τα ίδια τα δεδομένα, δηλαδή δεν είχε οριστεί εκ των προτέρων, και έγινε χειρωνακτικά και από ειδικούς. Αυτό οφείλεται στο γεγονός ότι ο κύριος στόχος του συστήματος κατηγοριοποίησης είναι να παρέχει συνδέσμους πλοήγησης στις σελίδες της Wikipedia μέσω μιας ιεραρχίας κατηγοριών. Κάτι τέτοιο βοηθά τους αναγνώστες, γνωρίζοντας τα χαρακτηριστικά που ορίζουν ένα θέμα, να εντοπίσουν και να περιηγηθούν γρήγορα σε ένα σύνολο σελίδων που ορίζονται με τα ίδια χαρακτηριστικά [II]. Καθοριστικό για την επιλογή της ως οντολογίας στο πλαίσιο της δουλειάς μας είναι και το γεγονός ότι πρόκειται για ηλεκτρονική εγκυκλοπαίδεια που διατίθεται ελεύθερα.

Συμπληρωματικά στα παραπάνω, αξίζει εδώ να αναφερθεί, το γεγονός ότι η αρχική μας σκέψη για την εξαγωγή και ανάθεση θέματος για καθεμιά από τις ιστοσελίδες υπό επεξεργασία ήταν άλλη, στην πορεία φάνηκε πως δεν ήταν η ιδανική, και ως εκ τούτου επαναπροσδιορίσαμε τη μεθοδολογία μας σε αυτό το κομμάτι και πήρε τη

μορφή που παρουσιάζεται αναλυτικά πιο πάνω. Συγκεκριμένα, σύμφωνα με την αρχική μας σκέψη, ο αλγόριθμος θεματικής κατηγοριοποίησης, κατά τη πρώτη φάση του, εκτός από τον εντοπισμό των στοιχείων «*P's anchorTitle*» και «*P's textTitle*», προέβλεπε και την εξαγωγή των λέξεων-κλειδιών εξαρχής και για όλες τις σελίδες. Με δεδομένα αυτά τα θεματικά στοιχεία, κατά τη δεύτερη Φάση (**ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 2**: Content-Based Classification, **Phase 2:** *Theme Detection*), προκειμένου να γίνει η ανάθεση του θέματος, ο αλγόριθμος έψαχνε, με τη βοήθεια της λεξικογραφικής οντολογίας WordNet τον πρώτο κοινό πρόγονο μεταξύ των ουσιαστικών των τριών αυτών θεματικών στοιχείων. Η σκέψη αυτή είχε προκύψει από την πεποίθηση ότι όσο πιο πολλά θεματικά στοιχεία ληφθούν υπόψη, τόσο πιο αντιπροσωπευτικό είναι το αποτέλεσμα. Ωστόσο, στην πράξη αποδείχθηκε πως δεν ήταν πρακτικός, ούτε ιδιαιτέρως αποτελεσματικός αυτός ο τρόπος ανάθεσης θέματος. Αρχικώς, έγινε αντιληπτό πως η εξαγωγή λέξεων-κλειδιών εξαρχής και για όλες τις ιστοσελίδες σημαίνει κατανάλωση χρόνου και υπολογιστικής ενέργειας, και σε συνδυασμό με το ότι η διαδικασία εντοπισμού κοινού προγόνου αποδείχθηκε επίσης χρονοβόρα και ενεργοβόρα, δίνοντας ένα πολύ πιο γενικό θέμα από αυτό που τελικώς λαμβάνεται από *«P's anchorTitle»* και *«P's textTitle»*. Έτσι θεωρήσαμε πως θα μπορούσαμε να επιλέξουμε μια πιο απλή και αποτελεσματική τακτική.

Σαν πρώτη εναλλακτική σκέψη, είχαμε την αναζήτηση, αντί κοινών προγόνων, κοινών όρων μεταξύ «*P's anchorTitle*» ή/και «*P's textTitle*»και λέξεων-κλειδιών. Όμως αυτό εξακολουθούσε να μην εξοικονομεί σημαντικό χρόνο και ενέργεια, έχοντας στην καλύτερη των περιπτώσεων, το ίδιο αποτέλεσμα με αυτό που θα έδιναν κοινοί όροι μεταξύ *P's anchorTitle* και *P's textTitle*, αν όχι πιο γενικό. Δεδομένου, λοιπόν, ότι τα δύο αυτά στοιχεία υπάρχουν σχεδόν σε όλες τις σελίδες, και μάλιστα πολύ συχνά ταυτίζονται και με το html element *<title> content*, αποδείχθηκε πως είναι υπέρ-αρκετά για την εξαγωγή θέματος. Οπότε, στην πράξη, δεν υπήρχε ανάγκη να γίνεται πάντα εξαγωγή λέξεων-κλειδιών και να καταναλώνονται χρόνος και ενέργεια, παρά μόνο όπου *P's anchorTitle* και *P's textTitle* δεν είναι αρκετά είτε επειδή δεν έχουν κοινούς όρους είτε επειδή δεν υπάρχουν και τα δύο αυτά στοιχεία σε κάποια σελίδα.

Τέλος, καθοριστικό ρόλο παίζει και το γεγονός ότι τα στοιχεία *P's anchorTitle* και *P's textTitle* αποτελούν έτοιμη πληροφορία, χωρίς να προϋποθέτουν τη χρήση κάποιου εργαλείου για τον εντοπισμό τους, με αποτέλεσμα να εξοικονομούνται και από αυτή τη σκοπιά χρόνος και ενέργεια, και χωρίς να υπάρχει πιθανότητα λάθους, όπως θα μπορούσε να υπάρχει με τη χρήση κάποιου εργαλείου.

4.3.2. Επανακατηγοριοποίηση Σελίδων Διαδικτύου με βάση τη Δυναμικότητά τους

Σκοπός των επιμέρους διαδικασιών που πραγματοποιούνται στο πλαίσιο τού δεύτερου αλγορίθμου (**ALGORITHM 2: Re-Classification based on Change Detection**), είναι ο διαχωρισμός των σελίδων σε αυτές που παραμένουν ίδιες και σε εκείνες που παρουσιάζουν αλλαγές με την πάροδο του χρόνου (έλεγχος δυναμικότητας δεδομένων), ο εντοπισμός των σημείων εκείνων όπου παρουσιάζονται οι αλλαγές, ο υπολογισμός τού βαθμού μεταβολής των ιστοσελίδων, και η λήψη απόφασης για την επανα-κατηγοριοποίησή τους ή όχι.

Αναλυτικότερα, προκειμένου να διαχωριστούν οι σελίδες σε αυτές που παραμένουν ίδιες και σε εκείνες που παρουσιάζουν αλλαγές με την πάροδο του χρόνου, απαιτείται ο έλεγχος της δυναμικότητας των σελίδων. Για το σκοπό αυτό, ο Αλγόριθμος 2 ελέγχει ανά τακτά χρονικά διαστήματα τα ζεύγη των ίδιων σελίδων σε διαφορετικές χρονικές στιγμές ως προς τα σημεία που εξετάστηκαν κατά τον Αλγόριθμο 1, ελέγχοντας παράλληλα την ημερομηνία τελευταίας ενημέρωσης της κάθε σελίδας, καθώς και το μέγεθός της σε κάθε μία από τις χρονικές στιγμές που υφίστανται επεξεργασία.

Πιο συγκεκριμένα, ο έλεγχος αυτός γίνεται σε δύο επίπεδα, θέματος και δομής. Κατά την πρώτη διαδικασία (**Procedure 1**: Re-Classification Decision based on Textual Changes), δηλαδή αυτής του ελέγχου της δυναμικότητας των σελίδων ως προς το θέμα τους, ο αλγόριθμος λαμβάνει ως δεδομένα εισόδου τις κατηγοριοποιημένες από τον Αλγόριθμο 1 (**Algorithm 1: MultiDimensional Page Classification**) σελίδες σε χρόνο *T* (*P(class, T)*), τις ίδιες σελίδες χωρίς να είναι κατηγοριοποιημένες σε χρόνο *T'(≠T)* (*P'(unclass, T')*), τις τιμές/το περιεχόμενο των κειμενικών στοιχείων που εξετάζονται

κατά τον Αλγόριθμο 1 για κάθε σελίδα σε χρόνο *T*, (*(E(t) ∈ P)*), (οπότε και γίνεται η αρχική κατηγοριοποίηση), και τις τιμές των ίδιων στοιχείων σε χρόνο *T'*, (*(E(t) ∈ P')*), (οπότε και πραγματοποιείται ο έλεγχος δυναμικότητας των σελίδων). Παράλληλα, ο αλγόριθμος χρειάζεται και μία μετρική ομοιότητας, (*sim(Pi, P'i)*), προκειμένου να συγκρίνει τις τιμές των στοιχείων που ορίζονται. Τέλος, ο αλγόριθμος χρειάζεται και δύο όρια, ((*t*), (*h*)), που ορίζουν τον μέγιστο επιτρεπτό από τον αλγόριθμο βαθμό ομοιότητας των σελίδων και τον ελάχιστο επιτρεπτό από τον αλγόριθμο βαθμό ομοιότητας των σελίδων αντίστοιχα. Στην εικόνα που ακολουθεί, Εικόνα 7, φαίνεται ο ψευδοκώδικας της πρώτης διαδικασίας του Αλγορίθμου 2.

ALGORITHM 2: RE-CLASSIFICATION BASED ON CHANGE DETECTION

```
1   Input: P(class, T), P'(unclass, T')
2     PROCEDURE 1: RE-CLASSIFICATION DECISION BASED ON TEXTUAL CHANGES
3     Input: (E(t) ∈ P), (E(t) ∈ P'), smlrtMetric, (m), (z)
4       for each pair of (Pi ∈ P(class,T), (P'i ∈ P'(unclass, T'))
5         compute sim(Pi, P'i)
6           if sim(Pi, P'i) ≥ m
7             tag P'i as thematically unchanged and classify P'i to the category of Pi
8             End
9           Else
10            tag P'i as thematically changed and
11              compare (E(t) ∈ P'i) with (E(t) ∈ Pi)
12              count ((E(t) ∈ P'i) ≠ (E(t) ∈ Pi))
13                if ((E(t) ∈ P'i) ≠ (E(t) ∈ Pi)) ≤ z
14                  go to Algorithm2Procedure2
15                  End
16                Else
17                  send P'i to Algorithm1Procedure2
18                  End
19      End
20    Output: thematically unchanged pages P' over time T'
```

Εικόνα 7: ALGORITHM 2: ReClassification based on Change Detection, **Procedure 1:** Re-Classification Decision based on ConTextual Changes

Με δεδομένα τα παραπάνω, ο Αλγόριθμος 2 κατά την πρώτη διαδικασία, εκτελεί διαδοχικά τις εξής ενέργειες: αξιοποιώντας τη μετρική ομοιότητας που έχει δοθεί ως δεδομένο εισόδου, συγκρίνει τις τιμές/το περιεχόμενο των θεματικών στοιχείων για τις ίδιες σελίδες σε διαφορετικές χρονικές στιγμές, και ελέγχει αν βαθμός ομοιότητάς τους είναι μικρότερος ή ίσος από το όριο (*t*) που έχει λάβει ως δεδομένο εισόδου. Αν είναι μεγαλύτερος από αυτό το όριο, η σελίδα θεωρείται στατική και άρα δεν χρειάζεται να κατηγοριοποιηθεί ξανά. Αν ο βαθμός ομοιότητας είναι μικρότερος από αυτό το όριο, ο αλγόριθμος καλείται να υπολογίσει σε ποιο βαθμό έχει υποστεί αλλαγές η σελίδα. Στην περίπτωση που ο βαθμός ομοιότητας είναι ίσος με το όριο (*t*)

που έχει δοθεί, ο αλγόριθμος δεν μπορεί να πάρει απόφαση. Έτσι, και κρατάει τη σελίδα με την αρχική της κατηγοριοποίηση, και τη στέλνει για υπολογισμό του βαθμού αλλαγής.

Για τον υπολογισμό του βαθμού αλλαγής, ο αλγόριθμος μετράει το πλήθος των θεματικών στοιχείων που διαφέρουν κατά τις χρονικές στιγμές που γίνεται ο έλεγχος δυναμικότητας, και δεδομένου του συνόλου των κειμενικών στοιχείων που ελέγχονται, υπολογίζει το ποσοστό αυτών που άλλαξαν. Αν αυτό είναι μεγαλύτερο από το όριο (*h*), που έχει συμπεριληφθεί στα δεδομένα εισόδου, η σελίδα στέλνεται ξανά για θεματική κατηγοριοποίηση, στη δεύτερη διαδικασία του πρώτου αλγορίθμου (**ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 2:** Content-Based Classification). Στην αντίθετη περίπτωση, αν δηλαδή είναι μικρότερος, η σελίδα θεωρείται στατική και παραμένει με την ήδη υπάρχουσα θεματική κατηγοριοποίηση. Ωστόσο, στέλνεται στο επόμενο βήμα, όπου γίνεται ο αντίστοιχος έλεγχος δυναμικότητας στα δομικά στοιχεία της. Στην περίπτωση που ο βαθμός ομοιότητας είναι ίσος με το όριο που έχει τεθεί, ο αλγόριθμος δεν μπορεί να πάρει απόφαση, κι έτσι κρατάει και την υπάρχουσα κατηγοριοποίηση, αλλά στέλνει την σελίδα για μια νέα στη δεύτερη διαδικασία του πρώτου αλγορίθμου (**ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 2:** Content-Base Classification). Ο λόγος για τον οποίο τα όρια δεν τίθενται με απόλυτη ισχύ, είναι για να λαμβάνονται όσο γίνεται πληρέστερα αποτελέσματα. Αποτέλεσμα της παραπάνω διαδικασίας είναι αρχικώς ο διαχωρισμός των σελίδων σε στατικές και σε μεταβαλλόμενες κατά το πέρασμα του χρόνου, και στη συνέχεια ο υπολογισμός του βαθμού αλλαγής των σελίδων, και άρα ο εντοπισμός εκείνων που χρήζουν θεματικής επανα-κατηγοριοποίησης.

Κατά τη δεύτερη διαδικασία του δεύτερου αλγορίθμου (**Procedure 2:** Re-Classification decision based on Structural Changes), δηλαδή του ελέγχου της δυναμικότητας των σελίδων ως προς τη δομή τους, στην ουσία γίνεται ό,τι και στην πρώτη διαδικασία, με τη διαφορά ότι ελέγχονται τα δομικά χαρακτηριστικά των σελίδων και όχι τα κειμενικά. Κατά συνέπεια ο δεύτερος αλγόριθμος κατά τη δεύτερη διαδικασία, λαμβάνει ως δεδομένα εισόδου τις κατηγοριοποιημένες από τον πρώτο αλγόριθμο

(**ALGORITHM1: Multi-Dimensional Page Classification**) σελίδες σε χρόνο *T*, *(P(class, T))*, τις ίδιες σελίδες χωρίς να είναι κατηγοριοποιημένες σε χρόνο *T'*, *(P'(unclass, T'))*, τις τιμές των δομικών στοιχείων που εξετάζονται κατά τον Αλγόριθμο 1 για κάθε σελίδα σε χρόνο *T*, *((E(s) ∈ P))*, οπότε και γίνεται η αρχική κατηγοριοποίηση, και τις τιμές των ίδιων στοιχείων σε χρόνο *T'*, *((E(s) ∈ P'))*, οπότε και πραγματοποιείται ο έλεγχος της δυναμικότητας των σελίδων. Τέλος, ο αλγόριθμος χρειάζεται και σε αυτό το σημείο ένα όριο *(h)*, που ορίζει τον επιτρεπτό από τον αλγόριθμο βαθμό ομοιότητας των σελίδων αντίστοιχα. Στην εικόνα που ακολουθεί, Εικόνα 8, φαίνεται ο ψευδοκώδικας της δεύτερης διαδικασίας του Αλγορίθμου 2 (**ALGORITHM 2: Re-Classification based on Change Detection, Procedure2:** Re-Classification Decision based on Structural Changes).

```
21   PROCEDURE2: RE-CLASSIFICATION DECISION BASED ON STRUCTURAL CHANGES
22   Input: (E(s) ∈ P), (E(s) ∈ P'), smlrtMetric, (z)
23       for each pair of (Pi ∈ P(class,T), (P'i ∈ P'(unclass, T'))
24           compare (E(s) ∈ Pi) with (E(s) ∈ P'i) and
25           count ((E(s) ∈ Pi) ≠ (E(s) ∈ P'i))
26               if ((E(s) ∈ Pi) ≠ (E(s) ∈ P'i)) ≤ z
27                   tag P'i as structurally unchanged and classify P'i to the category of Pi
28                   end
29               else
30                   send P'i to Algorithm1Procedure1
31                   End
32   Output: structurally unchanged pages P' over time T'
33 Output: P'(ReClass, T'), thematically unchanged pages P' over time T', structurally unchanged pages P' over time T'
```

Εικόνα 8: ALGORITHM 2: Re-Classification based on Change Detection, **Procedure 2:** Re-Classification based on Structural Changes

Κατά τη δεύτερη αυτή διαδικασία, ο αλγόριθμος συγκρίνει τις τιμές των δομικών στοιχείων που έχει εξετάσει κατά τον Αλγόριθμο 1 σε χρόνο *T*, με αυτές των ίδιων δομικών στοιχείων και για τις ίδιες σελίδες σε χρόνο *T'*. Στη συνέχεια, ο αλγόριθμος καλείται να μετρήσει το πλήθος των δομικών στοιχείων που εμφανίζουν διαφορές, και δεδομένου του συνόλου των δομικών στοιχείων που εξετάζονται, να υπολογίσει το βαθμό αλλαγής της κάθε σελίδας σε επίπεδο δομής. Ύστερα από τον υπολογισμό τού βαθμού αλλαγής, αν αυτός είναι μεγαλύτερος από το όριο (*h*), το οποίο επίσης συμπεριλαμβάνεται στα δεδομένα εισόδου, η σελίδα στέλνεται ξανά για δομική κατηγοριοποίηση, στην πρώτη διαδικασία του πρώτου αλγορίθμου (**ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 1:** Structure-Based Classification). Στην αντίθετη περίπτωση, αν δηλαδή είναι μικρότερος, η σελίδα

θεωρείται στατική και παραμένει με την ήδη υπάρχουσα δομική κατηγοριοποίηση. Στην περίπτωση που ο βαθμός ομοιότητας είναι ίσος με το όριο που έχει τεθεί, ο αλγόριθμος δεν μπορεί να πάρει απόφαση, κι έτσι κρατάει και την υπάρχουσα κατηγοριοποίηση, αλλά στέλνει την σελίδα για μια νέα στην πρώτη διαδικασία του πρώτου αλγορίθμου. Αποτέλεσμα της δεύτερης διαδικασίας (**ALGORITHM 2: Re-Classification based on Change Detection, Procedure 2:** Re-Classification Decision based on Structural Changes) είναι ο διαχωρισμός των δεδομένων υπό επεξεργασία, σε αυτά που παραμένουν στατικά ως προς τη δομή τους και σε εκείνα που αλλάζουν ως προς τη δομή τους, και μάλιστα σε βαθμό τέτοιο που επιβάλλεται η δομική επανακατηγοριοποίησή τους.

Κατά το σχεδιασμό του εν λόγω αλγορίθμου, αρχικώς θεωρήσαμε πως θα ήταν πολύ χρήσιμο ο δεύτερος να λάβει υπόψη του και το στοιχείο της τελευταίας επικαιροποίησης της σελίδας (*last update*), έτσι όπως αυτό φαίνεται μέσα στον html κώδικά της, καθώς και αυτό του μεγέθους της ίδιας σελίδας σε διαφορετικές στιγμές. Ωστόσο, στην πράξη, διαπιστώθηκε πως υπάρχουν αρκετοί εναλλακτικοί τρόποι για την καταχώριση της πληροφορίας που αφορά στην ημερομηνία της πιο πρόσφατης έκδοσής της, όπως είναι τα διάφορα σχετικά html elements ή η καθαρή ημερομηνία μέσα στο σώμα της σελίδας που βλέπει ο χρήστης. Επίσης, η ημερομηνία δημιουργίας της πιο πρόσφατης έκδοσης ενός ιστοτόπου, μπορεί να διαφέρει από αυτό μιας σελίδας, ή/και η πληροφορία να αφορά τις πλευρικές στήλες της σελίδας και όχι το κυρίως σώμα της. Αντίστοιχα, το ζήτημα του συνυπολογισμού της αλλαγής μεγέθους της σελίδας, στην πράξη φάνηκε πολύ γενικό, μιας και το μέγεθος μιας σελίδας μπορεί να αλλάζει επειδή αλλάζει η γραμματοσειρά, το χρώμα ή η ανάλυση μιας εικόνας, γεγονός που δεν συνιστά θεματική ή δομική αλλαγή.

Ανακεφαλαιώνοντας, συνολικά στο πλαίσιο του δεύτερου αλγορίθμου βγαίνει αρχικώς το συμπέρασμα για το αν οι σελίδες αλλάζουν είτε ως προς το περιεχόμενο είτε/και ως προς τη δομή τους, και ύστερα σε ποιο βαθμό. Εκείνες που δεν παρουσιάζουν κάποιου είδους αλλαγή σε κάποια από τις χρονικές στιγμές κατά τις οποίες εξετάζονται, χαρακτηρίζονται στατικές, και η πρώτη κατηγοριοποίηση που έγινε θεωρείται οριστική. Ενώ αυτές που παρουσιάζουν αλλαγές, οι οποίες είναι

εντός των επιτρεπτών από τον αλγόριθμο ορίων, υφίστανται εκ νέου κατηγοριοποίηση θεματική ή/και δομική.

4.3.3. Βελτιστοποίηση Επανακατηγοριοποίησης Σελίδων Διαδικτύου με βάση τη Συχνότητα Αλλαγής

Σκοπός του τρίτου και τελευταίου για την προτεινόμενη μεθοδολογία (**ALGORITHM 3: Optimized Re-Classification based on Change's Frequency Detection**), είναι να βελτιώσει τη λειτουργία τού δεύτερου αλγορίθμου (**ALGORITHM 2: Re-Classification based on Textual Changes**). Όπως είδαμε στην προηγούμενη ενότητα, με την εφαρμογή του τελευταίου, ελέγχεται η δυναμικότητα των σελίδων, και επιλέγονται εκείνες που χρήζουν επανα-κατηγοριοποίησης. Ο Αλγόριθμος 3 μπορεί να βελτιώσει τη διαδικασία αυτή, υπολογίζοντας το ρυθμό αλλαγής των σελίδων, δηλαδή πόσο συχνά μεταβάλλονται ως προς το περιεχόμενο ή/και ως προς τη δομή τους. Ένας τέτοιος υπολογισμός, μπορεί να βελτιώσει τη λειτουργία του Αλγορίθμου 2, καθώς ελέγχοντας το ρυθμό μεταβολής των σελίδων, μπορεί να αξιολογηθεί ποιες σελίδες «έχει νόημα», πρακτικά και ουσιαστικά, να σταλούν στον Αλγόριθμο 2 για επανα-κατηγοριοποίηση και ποιες όχι, με αποτέλεσμα να εξοικονομούμε χρόνο και υπολογιστικό κόστος.

Η διάκριση αυτή βασίζεται στο γεγονός ότι σε περιπτώσεις όπου οι σελίδες αλλάζουν σπάνια, δηλαδή ο ρυθμός μεταβολής τους είναι ιδιαιτέρως χαμηλός, είναι σπατάλη υπολογιστικής δύναμης να κατηγοριοποιούνται ξανά και ξανά, αφού το αποτέλεσμα της νέας κατηγοριοποίησης θα είναι τις περισσότερες φορές όμοιο με αυτό της αρχικής. Από την άλλη, σε περιπτώσεις όπου οι σελίδες εμφανίζουν υψηλό ρυθμό μεταβολής, επίσης είναι σπατάλη υπολογιστικών πόρων και δύναμης να προσπαθούμε να τις κατηγοριοποιούμε κάθε φορά που αλλάζουν, δεδομένου ότι ο αλγόριθμος μπορεί και να μην προλαβαίνει να ολοκληρωθεί πριν εκείνες αλλάξουν εκ νέου. Αυτό σημαίνει ότι μία σελίδα μπορεί να σταλεί για επανα-κατηγοριοποίηση, δεδομένου ότι έχει αλλάξει, αλλά μέχρι να ολοκληρωθεί η διαδικασία της επανα-κατηγοριοποίησης, να έχει αλλάξει ξανά. Οπότε, το αποτέλεσμα της επανα-κατηγοριοποίησης ουσιαστικά δεν ισχύει, αφού, τη στιγμή που μας δίνεται, δεν αφορά την τρέχουσα έκδοση της εκάστοτε σελίδας, αλλά προγενέστερη. Κατά

συνέπεια, βασική λειτουργία του Αλγορίθμου 3 είναι ο έλεγχος για το πόσο συχνά παρατηρείται κάποια αλλαγή στις σελίδες που εξετάζονται, με σκοπό τη βελτιστοποίηση της διαδικασίας της επανα-κατηγοριοποίησης.

Από τα παραπάνω, γίνεται σαφής ο πολύ σημαντικός ρόλος που παίζει το χρονόμετρο κατά την εφαρμογή του Αλγορίθμου 3, το οποίο, όπως εξηγείται στη συνέχεια, συμπεριλαμβάνεται στα δεδομένα εισόδου και λειτουργεί βάσει ενός ορισμένου «κανόνα». Αφετηρία για το σχεδιασμό του Αλγορίθμου 3, και ιδίως το κομμάτι της ενσωμάτωσης ενός χρονομέτρου, και ο ρόλος που αυτό παίζει στο πλαίσιο του εν λόγω Αλγορίθμου είναι η δουλειά [50]. Στην εικόνα που ακολουθεί, φαίνεται ο ψευδοκώδικας τού Αλγορίθμου3.

ALGORITHM 3: OPTIMIZED RE-CLASSIFICATION BASED ON CHANGE'S FREQUENCY DETECTION

```
1   Input: (P(class, T)), ((E(t) U E(s)) ∈ P(class, T)), P' ⊆ (P'(re-class, T'), ((E(t) U E(s)) ∈ P'(reClass,T')),
    MaxFreqChange, MinFreqChange, Timer
2       when Algorithm 2 initializes, record Ts
3       for every pair of ((Pi ∈ P(class, T), (P'i ∈ P'(re-class, T'))
4           set Timer
5             while ((E(t) U E(s)) ∈ Pi(class, T)) ≠ ((E(t) U E(s)) ∈ P'i(re-class,T')), record Ts
6               if Ts ≥ MaxFreqChange
7                 tag P'i as HighlyChanging Page and keep it in a secondary Index
8                 End
9               else
10                if Ts ≤ MinFreqChange
11                  tag P'i as RarelyChanging Page and keep it in a secondary Index
12                  end
13                else
14                  tag P'i as RegularlyChanging Page and send it to Algorithm2
15                  end
16      End
17  Output: Selection of Pages that need periodical Re-Classification
```

Εικόνα 9: ALGORITHM 3: Optimized ReClassification based on Change's Frequency Detection

Αναλυτικά, τα δεδομένα εισόδου που χρειάζεται ο Αλγόριθμος 3 είναι οι σελίδες που κατηγοριοποιούνται σε χρόνο *T*, (*P(class, T)*), τα κειμενικής φύσης και τα δομικής φύσης στοιχεία που χαρακτηρίζουν τις κατηγοριοποιημένες σε χρόνο *T* σελίδες και τα οποία μελετώνται κατά τον Αλγόριθμο 1, *(E(t) U E(s)) ∈P(class, T))*, οι σελίδες εκείνες που επανα-κατηγοριοποιήθηκαν από τον Αλγόριθμο 2 σε χρόνο *T'*, *(P' ⊆ (P'(re-class, T'))*, τα κειμενικής φύσης και τα δομικής φύσης στοιχεία που ανήκουν στις τελευταίες, (*(E(t) U E(s)) ∈ P'(re-class,T')*), η μέγιστη επιτρεπτή από τον Αλγόριθμο συχνότητα εμφάνισης αλλαγών στις σελίδες (*MaxFreqChange*), η ελάχιστη επιτρεπτή από τον Αλγόριθμο συχνότητα εμφάνισης αλλαγών στις σελίδες (*MinFreqChange*), το

χρονόμετρο (*Timer*) το οποίο υπολογίζει το χρόνο βάσει ενός ορισμένου τύπου που εξηγείται στη συνέχεια.

Ο ρόλος του χρονομέτρου είναι να μετράει το χρόνο, ώστε η λειτουργία του Αλγορίθμου 3 να ξεκινάει σε προγραμματισμένες χρονικές στιγμές που εμείς ορίζουμε, προκειμένου να εξυπηρετηθεί ο σκοπός μας. Με άλλα λόγια, μέσω του χρονομέτρου καθορίζονται οι χρονικές στιγμές κατά τις οποίες ο Αλγόριθμος 3 ελέγχει τις σελίδες για ενδεχόμενες αλλαγές. Ο κανόνας βάσει του οποίου προγραμματίζεται είναι ο εξής: αν *t* η χρονική στιγμή που εξετάστηκε για πρώτη φορά μια σελίδα, το χρονόμετρο θα υπολογίζει την κάθε *ti* χρονική στιγμή που θέλουμε να εξεταστεί ξανά και ξανά η ίδια σελίδα βάσει του τύπου *ti=(ti-1)*σταθερά*. Δηλαδή, κάθε χρονική στιγμή επανεξέτασης θα προκύπτει από το γινόμενο της αμέσως προηγούμενής της με μια σταθερά που θα είναι ορισμένη.

Με δεδομένα τα παραπάνω, πρώτη ενέργεια κατά την εκτέλεση του Αλγορίθμου 3, είναι η καταγραφή της χρονικής στιγμής (*Ts*) που ξεκινάει να λειτουργεί. Στη συνέχεια, και για κάθε ζεύγος (*(E(t) U E(s) ∈ P(class, T)*, (*E(t) U E(s) ∈ P'(re-class, T')*), τίθεται σε λειτουργία το χρονόμετρο. Ύστερα, και κάθε φορά που ο αλγόριθμος εντοπίζει τη συνθήκη (*E(t) U E(s) ∈ P(class, T) ≠ (E(t) U E(s) ∈ P'(re-class, T')*), καταγράφει τη χρονική στιγμή (*Ts*). Τέλος, με βάση το πλήθος των καταγεγραμμένων χρονικών στιγμών (*Ts*) που παρατηρείται κάποια διαφοροποίηση για την ίδια σελίδα, λαμβάνεται η απόφαση για την επανα-κατηγοριοποίησή της ή μη.

Συγκεκριμένα, αν το πλήθος των καταγεγραμμένων χρονικών στιγμών, ως στιγμών εντοπισμού κάποιας αλλαγής μεταξύ των διαφορετικών χρονικά στιγμιοτύπων της ίδιας σελίδας, είναι περισσότερες από τη μέγιστη επιτρεπτή από τον αλγόριθμο συχνότητα εμφάνισης αλλαγών (*MaxFreqChange*), τότε η σελίδα χαρακτηρίζεται ως υψηλής συχνότητας μεταβαλλόμενη σελίδα (*HighlyChanging page*) και δεν στέλνεται στον Αλγόριθμο 2 για επανα-κατηγοριοποίηση. Αν το πλήθος των καταγεγραμμένων χρονικών στιγμών, ως στιγμών εντοπισμού κάποιας αλλαγής μεταξύ των διαφορετικών χρονικά στιγμιοτύπων της ίδιας σελίδας, είναι λιγότερες από την ελάχιστη επιτρεπτή από τον αλγόριθμο συχνότητα εμφάνισης αλλαγών (*MinFreqChange*), τότε η σελίδα χαρακτηρίζεται ως χαμηλής συχνότητας

μεταβαλλόμενη σελίδα (*RarelyChanging page*) και επίσης δεν στέλνεται στο Αλγόριθμο 2 για επανα-κατηγοριοποίηση. Σε κάθε άλλη περίπτωση, όπου το πλήθος των καταγεγραμμένων χρονικών στιγμών, ως στιγμών εντοπισμού κάποιας αλλαγής μεταξύ των διαφορετικών χρονικά στιγμιοτύπων της ίδιας σελίδας, βρίσκεται εντός των παραπάνω ορίων, τότε η σελίδα χαρακτηρίζεται ως συνηθισμένης συχνότητας μεταβαλλόμενη σελίδα (*RegularlyChanging Page*) και στέλνεται στον Αλγόριθμο 2 για επανα-κατηγοριοποίηση.

Αποτέλεσμα της διαδικασίας αυτής είναι η επιλογή των σελίδων εκείνων που - κατόπιν παρατήρησης και δεδομένων ορισμένων ορίων, τα οποία μπορούν να αλλάζουν – χρειάζεται να επανα-κατηγοριοποιούνται ανά τακτά χρονικά διαστήματα.

4.4. Σύνοψη Μεθοδολογίας

Στις παραπάνω Ενότητες παρουσιάζονται αναλυτικά τα βήματα κάθε διαδικασίας που περιλαμβάνει η προτεινόμενη μεθοδολογία, ενώ στη συνέχεια μελετώνται η απόδοση τού προτεινόμενου αλγορίθμου και τα συμπεράσματα που προκύπτουν από την πειραματική του εφαρμογή.

Όπως αναφέρεται και στις προηγούμενες ενότητες, κατά τον καθορισμό των επιμέρους διαδικασιών και διεργασιών τού αλγορίθμου μας, καλούμαστε να κάνουμε κάποιες σχεδιαστικές επιλογές. Οι επιλογές αυτές αφορούν τον αποκλεισμό ορισμένων δεδομένων ή/και των υπολογισμό άλλων. Η διαδικασία αυτή, καθώς και οι τελικές μας αποφάσεις στηρίζονται τόσο στην υπάρχουσα βιβλιογραφία, όσο και στην δική μας παρατήρηση κατά τις πρώτες πειραματικές δοκιμές. Τέλος, σημαντικό ρόλο παίζουν οι πόροι και τα εργαλεία που έχουμε στη διάθεσή μας, αλλά και το σύνολο των δεδομένων πάνω στο οποίο πειραματιζόμαστε. Σκοπός μας σε κάθε περίπτωση ήταν η αξιοποίηση της υπάρχουσας γνώσης, με στόχο την εξέλιξή της.

ΚΕΦΑΛΑΙΟ 5: Πειραματική Αξιολόγηση

5.1. Εισαγωγή

Σε αυτό το Κεφάλαιο γίνεται η περιγραφή της πειραματικής δοκιμής και μελέτης της προτεινόμενης μεθοδολογίας. Με άλλα λόγια, περιγράφονται αναλυτικά η πειραματική της δοκιμή, τα αποτελέσματα που λαμβάνονται από αυτήν, καθώς και η συγκριτική παρουσίαση των αποτελεσμάτων αυτών σε σχέση με αυτά που λαμβάνονται από την εφαρμογή ενός παραδοσιακού κατηγοριοποιητή. Σκοπός μας μέσα από τη διαδικασία της πειραματικής δοκιμής και μελέτης, είναι η αξιολόγηση του προτεινόμενου αλγόριθμου κατηγοριοποίησης ως προς την αποτελεσματικότητά του, αλλά και ως προς την ορθότητα των αποτελεσμάτων του.

Πιο συγκεκριμένα, για την πραγματοποίηση της πειραματικής αξιολόγησης στο πλαίσιο της παρούσας διατριβής, αρχικώς δοκιμάζεται η προτεινόμενη μεθοδολογία σε ένα σύνολο δεδομένων (ιστοσελίδες διαδικτύου) μικρής κλίμακας. Αυτή η διαδικασία επιτρέπει τον έλεγχο τόσο της απόδοσης της μεθοδολογίας μας, όσο και τις ενδεχόμενες αδυναμίες της. Στη συνέχεια, και έχοντας καταγράψει τα αποτελέσματα της πειραματικής δοκιμής, ελέγχεται η ορθότητά τους αξιοποιώντας μετρικές αξιολόγησης ποιότητας αλγορίθμων. Πρόκειται για τις μετρικές ακρίβειας και ανάκλησης, που χρησιμοποιούνται ευρέως για την αξιολόγηση της ποιότητας ενός συστήματος ταξινόμησης, και οι οποίες εξηγούνται περισσότερο στη συνέχεια.

Στις ενότητες που ακολουθούν, περιγράφονται αναλυτικά η πειραματική δοκιμή της προτεινόμενης μεθοδολογίας (5.2. Περιγραφή Πειραματικής Δοκιμής) και τα αποτελέσματα που λαμβάνονται (5.3. Πειραματικά Αποτελέσματα) από αυτήν. Στη συνέχεια, παρουσιάζεται αναλυτικά ο τρόπος που επιλέγουμε να αξιολογήσουμε την επάρκεια της απόδοσής της βάσει καθιερωμένων μετρικών (5.4. Μετρικές Αξιολόγησης). Τέλος, περιγράφεται το πλαίσιο διεξαγωγής της συγκριτικής μελέτης (5.5. Συγκριτική Μελέτη), καθώς και τα αποτελέσματα αυτής (5.6. Αποτελέσματα συγκριτικής πειραματικής μελέτης).

5.2. Περιγραφή Πειραματικής Δοκιμής

Όπως αναφέρεται και στην εισαγωγή του Κεφαλαίου (5.1. Εισαγωγή) η πειραματική δοκιμή της προτεινόμενης μεθοδολογίας γίνεται σε ένα μικρής κλίμακας σύνολο δεδομένων, σε 2.330 τυχαίες και ποικίλες, ως προς τη μορφή και το περιεχόμενο, σελίδες διαδικτύου. Επιλέγεται αυτός ο αριθμός, καθώς θεωρείται ο μέσος όρος των αποθηκευμένων σελίδων διαδικτύου ενός έμπειρου χρήστη. Προκειμένου να συλλεχθεί το δείγμα δεδομένων στο οποίο εφαρμόζεται η προτεινόμενη μεθοδολογία, ζητείται από 10 εθελοντές, έμπειρους χρήστες του διαδικτύου, να μας κοινοποιήσουν τους σελιδοδείκτες που έχουν αποθηκευμένους στον περιηγητή τους. Με αυτόν τον τρόπο, συλλέγεται ένα αντιπροσωπευτικό σύνολο τυχαίων και ποικίλων ως προς τη θεματολογία και την τυπολογία τους δεδομένα (όπου, δεδομένα, εννοούμε τις σελίδες διαδικτύου).

Παράλληλα με τη συλλογή των δεδομένων, πληροφορούμε τους εθελοντές σχετικά με το αντικείμενο και το σκοπό της έρευνας στο πλαίσιο της παρούσας διατριβής, και τους ζητάμε να χαρακτηρίσουν κάθε έναν από τους σελιδοδείκτες που μοιράζονται μαζί μας ως προς τον τύπο τους, με δεδομένο το σκοπό που εξυπηρετούν (πληροφοριακή σελίδα, σελίδα πλοήγησης ή σελίδα διάδρασης). Προκειμένου να εξοικειωθούν οι εθελοντές με τους τρεις τύπους σελίδων διαδικτύου, τους παρέχονται σύντομες οδηγίες σχετικά με τον ορισμό κάθε τύπου και τους «εκπαιδεύουμε» δίνοντάς τους αρκετά παραδείγματα. Το ζητούμενο είναι να χαρακτηρισθεί κάθε σελίδα που μας παρέχουν με έναν και μοναδικό τύπο. Γι’ αυτό, στις περιπτώσεις που δεν είναι βέβαιοι για τον τύπο της σελίδας, τους ζητείται να αφαιρούν τη σελίδα από το δείγμα που θα μας παρέχουν. Συμπληρωματικά στα παραπάνω, ζητείται από τους εθελοντές να ορίσουν το θέμα της κάθε σελίδας όπως το αντιλαμβάνονται οι ίδιοι, και να το εκφράσουν με δικά τους λόγια. Οι εθελοντές έχουν την ελευθερία να εκφράσουν το θέμα κάθε σελίδας με όσες λέξεις-κλειδιά επιθυμούν, υπογραμμίζοντας εκείνο που θεωρούν πιο αντιπροσωπευτικό.

Με αυτόν τον τρόπο, συλλέγονται 2.330 τυχαίες και διαφορετικές σελίδες διαδικτύου, χαρακτηρισμένες από τους εθελοντές ως προς τον τύπο και το θέμα τους.

Από το σύνολο των πειραματικών δεδομένων αποκλείουμε εξαρχής ορισμένα είδη σελίδων που αποτελούν ιδιαίτερες περιπτώσεις, όπως είναι οι ιστότοποι κοινωνικής δικτύωσης και ο ιστότοπος *youTube*. Οι δύο αυτοί τύποι σελίδων διαδικτύου, αποτελούν από μόνες τους ξεχωριστές κατηγορίες, αφού έχουν ξεχωριστά ιδιαίτερα χαρακτηριστικά (π.χ. στο youTube συναντάμε κυρίως οπτικοακουστικό υλικό, στα μέσα κοινωνικής δικτύωσης συναντάμε συνεχή ροή πληροφοριών και εμπλοκής μέσω «αντιδράσεων» των χρηστών σε πραγματικό χρόνο), η μελέτη των οποίων δεν συμπεριλαμβάνεται στους σκοπούς της παρούσας διατριβής.

Σχετικά με τη διαδικασία της καθαυτής πειραματικής δοκιμής της προτεινόμενης μεθοδολογίας, χρειάζεται να διευκρινιστεί πως, πριν από αυτήν, τα δεδομένα προς επεξεργασία και κατηγοριοποίηση «καθαρίζονται» από κάθε επισημείωση που είχαν κάνει η εθελοντές που μας παρείχαν τους σελιδοδείκτες τους. Στη συνέχεια, επεξεργαζόμαστε το δείγμα δεδομένων ακολουθώντας πιστά τα βήματα των αλγορίθμων μας και καταγράφουμε τα αποτελέσματα. Τέλος, ποσοτικοποιούμε την ακρίβεια των αποτελεσμάτων του αλγορίθμου κατηγοριοποίησης, χρησιμοποιώντας τους τύπους της ανάκλησης κ της ακρίβειας.

5.2.1. Πολυδιάστατη Κατηγοριοποίηση Σελίδων Διαδικτύου (ALGORITHM1: Multi-Dimensional Page Classification)

Ο πρώτος αλγόριθμος που καλούμαστε να εκτελέσουμε στο πλαίσιο της πειραματικής δοκιμής είναι αυτός της πολυδιάστατης κατηγοριοποίησης των σελίδων διαδικτύου (**ALGORITHM 1: Multi-Dimensional Page Classification**), μέσω του οποίου κατηγοριοποιούνται οι σελίδες διαδικτύου ως προς τη δομή και το θέμα τους. Όπως φαίνεται και από την αναλυτική περιγραφή της προτεινόμενης μεθοδολογίας, ΚΕΦΑΛΑΙΟ 4: Προτεινόμενη Μεθοδολογία, ο αλγόριθμος για την πολυδιάστατη κατηγοριοποίηση των σελίδων απαρτίζεται από δύο επιμέρους, ξεχωριστές και ανεξάρτητες η μία από την άλλη διαδικασίες (**Procedure 1:** Structure-Based Classification και **Procedure 2**: Content-Based Classification), που μπορούν να εκτελεστούν παράλληλα.

Ξεκινώντας την περιγραφή από τη διαδικασία της δομικής κατηγοριοποίησης, κατά την πρώτη φάση (**Phase 1:** *Page Type Recognition*) διαχωρίζονται οι σελίδες ως προς τον τύπο τους σε σελίδες *διάδρασης*, *πλοήγησης* και *πληροφοριακές*. Με βάση τον προτεινόμενο αλγόριθμο, κατά το πρώτο βήμα, ελέγχεται αν στο κυρίως σώμα της κάθε σελίδας, υπάρχει τουλάχιστον ένας όρος διάδρασης. Με αυτόν τον τρόπο, ήδη από το πρώτο βήμα εντοπίζονται οι σελίδες διάδρασης. Οι όροι διάδρασης, είναι οι όροι εκείνοι (υπο-)δηλώνουν την ύπαρξη κάποιας ενέργειας διάδρασης με τον χρήστη. Με βάση τους όρους που συναντάμε στη βιβλιογραφία [38], όσους γνωρίζουμε από την πείρα μας ως χρήστες τού διαδικτύου, αλλά και άλλους που συναντάμε στις σελίδες του πειραματικού μας δείγματος, φτιάχνουμε τον *Πίνακα όρων Διάδρασης* (***T(trans):*** *table with transactional terms*) (Πίνακας 1). Έτσι, κάθε φορά που στο κυρίως σώμα του περιεχομένου μιας σελίδας, εντοπίζεται ένας ή περισσότεροι από αυτούς τους όρους διάδρασης, η σελίδα χαρακτηρίζεται αναλόγως (δηλαδή, σελίδα διάδρασης), με αποτέλεσμα αυτές να ξεχωρίζουν αμέσως.

Serial Number	Transactional Terms
1	Shop Now
2	(Add to) Bag
3	(Add to) Basket
4	(Add to) Cart
5	Book
6	Book a table
7	Book Now
8	Book return
9	Book Single
10	Buy
11	Buy from App Store
12	Buy from RouteBuddy
13	Buy from Splash Maps
14	Buy from the National Trails shop
15	Buy Now
16	Buy Online
17	Download
18	e-Gift Card
19	Find a table
20	Find Tickets

Πίνακας 1: Ενδεικτικός πίνακας όρων Διάδρασης [**T(trans):** table with transactional terms (**t(trans)**)].

Στη συνέχεια, ακολουθώντας το αμέσως επόμενο βήμα του Αλγορίθμου 1, και προκειμένου να διαχωριστούν οι σελίδες πλοήγησης από τις πληροφοριακές,

χρειάζεται ο υπολογισμός της αναλογίας κειμένου-συνδέσμων για το περιεχόμενο κάθε σελίδας. Χρήσιμο εργαλείο, το οποίο και αξιοποιείται σε αυτό το βήμα, αποδεικνύεται το Text to Link Analyzer [III], το οποίο διατίθεται ελεύθερα στο διαδίκτυο. Μέσα από αυτό, εισάγοντας στο περιβάλλον χρήστη του εργαλείου τον σύνδεσμο που θέλουμε να επεξεργαστούμε, λαμβάνουμε τα εξής στοιχεία για τη σελίδα που αντιπροσωπεύει: το ποσοστό περιεχόμενου στη σελίδα κειμένου, το ποσοστό περιεχόμενων στη σελίδα συνδέσμων, και την αναλογία κειμένου/συνδέσμων. Πρόκειται για ένα εργαλείο αξιόπιστο, που το ενδεχόμενο ποσοστό λάθους του δεν επηρεάζει την αποτελεσματικότητα του αλγορίθμου μας. Η μοναδική αδυναμία που εμείς ως χρήστες εντοπίζουμε, είναι ότι δεν διαχωρίζει -τουλάχιστον εμφανώς- το κυρίως σώμα του περιεχομένου μιας σελίδας, από αυτό που μπορεί να βρίσκεται σε σταθερό πλαίσιο σε κάποιο από τα άκρα της (άνω/κάτω/πλαϊνά άκρα). Ύστερα από σχετικό εμπειρικό έλεγχο, και δεδομένης της ταχύτητας απόκρισής του και του τρόπου παρουσίασης των αποτελεσμάτων, διαπιστώνουμε πως η αδυναμία αυτή δεν συνιστά εμπόδιο στη χρήση του προκειμένου να εξυπηρετηθεί ο σκοπός της παρούσας πειραματικής δοκιμής.

Σύμφωνα με τον προτεινόμενο αλγόριθμο, και έχοντας υπολογίσει την αναλογία κειμένου/συνδέσμων για κάθε σελίδα, χρειάζεται να τεθεί το όριο για τον διαχωρισμό των σελίδων πλοήγησης από τις πληροφοριακές. Αρχικώς επιλέγεται ένα αρκετά ευρύ και ασφαλές όριο, ώστε να αποφευχθεί στο μέγιστο δυνατό βαθμό το ενδεχόμενο λανθασμένης απόφασης σχετικά με τον τύπο της κάθε σελίδας. Συγκεκριμένα, τίθεται το όριο τουλάχιστον *70% κείμενο* και *30% σύνδεσμοι* για να χαρακτηριστεί μια σελίδα πληροφοριακή και στην αντίθετη περίπτωση, όπου το ποσοστό αυτό είναι υπέρ των συνδέσμων χαρακτηρίζεται πλοήγησης. Δηλαδή, 70% ποσοστό κειμένου κατ' ελάχιστο και 30% μέγιστο ποσοστό συνδέσμων για να χαρακτηριστεί πληροφοριακή ή 70% ποσοστό συνδέσμων κατ' ελάχιστο και 30% μέγιστο ποσοστό κειμένου για να χαρακτηριστεί πλοήγησης. Βλέποντας τα αποτελέσματα, με δεδομένο αυτό το αρκετά αυστηρό όριο, διαπιστώνεται πως είναι αρκετά υψηλό το ποσοστό των σελίδων για τις οποίες ο προτεινόμενος αλγόριθμος δεν παίρνει απόφαση για τον τύπο τους. Αυτό οφείλεται στο γεγονός ότι είναι πολλές οι σελίδες που η αναλογία κειμένου/συνδέσμων του περιεχομένου τους βρίσκονται

εντός αυτών των ορίων, με αποτέλεσμα να μένουν χωρίς ετικέτα για τον τύπο τους. Προκειμένου να μειωθεί ο αριθμός των σελίδων τις οποίες ο προτεινόμενος αλγόριθμος δεν μπορούσε να χαρακτηρίσει ως προς το είδος τους, ορίζεται πιο ευρύ όριο. Συγκεκριμένα, το νέο όριο είναι τουλάχιστον *60% κείμενο* και *40% σύνδεσμοι* για να χαρακτηριστεί μια σελίδα πληροφοριακή και στην αντίθετη περίπτωση, χαρακτηρίζεται πλοήγησης. Αξίζει εδώ να υπογραμμίσουμε, το γεγονός ότι αυτό είναι το όριο που προκύπτει από την πειραματική εφαρμογή που γίνεται στο συγκεκριμένο σύνολο δεδομένων. Αυτό σημαίνει ότι σε ένα άλλο σύνολο δεδομένων, το όριο μπορεί να είναι διαφορετικό. Κατά συνέπεια, το όριο του αλγορίθμου δεν είναι απόλυτο, και μπορεί να διαφοροποιείται ανάλογα με το σύνολο των σελίδων.

Αποτέλεσμα των παραπάνω ενεργειών είναι ο διαχωρισμός των σελίδων με βάση το είδος τους σε διάδρασης, πλοήγησης και πληροφοριακές. Στη συνέχεια, βάσει του είδους της κάθε σελίδας, πραγματοποιείται η βαθύτερη δομική κατηγοριοποίηση. Οι σελίδες διάδρασης κατηγοριοποιούνται σε *free* και *not-free*, με βάση το αν απαιτείται οικονομική συναλλαγή με τον χρήστη προκειμένου να ολοκληρωθεί η ενέργεια. Ενώ οι σελίδες πλοήγησης κατηγοριοποιούνται σε *WebPage* και *HomePage*, και στη συνέχεια οι πρώτες επισημειώνονται με βάση το βάθος το οποίο βρίσκονται στον ιστότοπο που τις φιλοξενεί, ενώ οι δεύτερες κατηγοριοποιούνται με τη σειρά τους με βάση τον *τομέα* (*domain*) στο οποίο ανήκουν.

Για τις σελίδες διάδρασης, σύμφωνα με τον προτεινόμενο αλγόριθμο, κατά το πρώτο βήμα της δεύτερης φάσης της δομικής κατηγοριοποίησης (**Phase 2:** Layered Page Classification), επιβάλλεται να ελεγχθεί κάτω από ποια/ες από τις κατηγορίες του σχετικού πίνακα (**Πίνακας 2:** Πίνακας όρων συσχέτισης (T(corr): **Table of correlation**)), είναι οργανωμένοι οι περισσότεροι από τους όρους διάδρασης που εμφανίζονται στην εκάστοτε σελίδα. Με αυτόν τον τρόπο, εντοπίζεται το είδος της συναλλαγής για την οποία δημιουργήθηκε η κάθε σελίδα, και ετικετοποιούμε την καθεμιά αναλόγως. Για την σύσταση του εν λόγω πίνακα (Πίνακας 2), βασιζόμαστε στη βιβλιογραφία [39], στη σημασία των λέξεων και στην πείρα μας ως χρηστών του διαδικτύου.

Serial Number	Banking	Booking	Download	E-commerce	Entertainment	Software
1	i-bank pass	Book a table	Download	Shop Now	Download	Download
2	i-bank pay	Book Now	Free trial	(Add to) Bag	Find a table	Free trial
3		Book return	Games	(Add to) Basket	Find Tickets	Games
4		Book Single	Take the test	(Add to) Cart	Free trial	Sale
5		Find a table		(Add to) compare	Games	Take the test
6		Find Tickets		Best Seller	Listen now	
7		Make a booking		Book	Make a booking	
8		Special offers		Book a table	Play	
9				Book Now	Take the test	
10				Book return	Video	

Πίνακας 2: Πίνακας όρων συσχέτισης (T(corr): Table of correlation)

Ακολουθώντας την ίδια τακτική με αυτή που ακολουθείται για την κατάρτιση του πίνακα με τους όρους διάδρασης, προκύπτει ο πίνακας με τους όρους που (υπο-)δηλώνουν οικονομική συναλλαγή, **Πίνακας 3:** Πίνακας όρων συναλλαγής**.**

Serial Number	Payment Terms
1	Shop Now
2	(Add to) Bag
3	(Add to) Basket
4	(Add to) Cart
5	(Add to) compare
6	Best Seller
7	Book
8	Book a table
9	Book Now
10	Book return
11	Book Single
12	Buy
13	Buy from App Store
14	Buy from RouteBuddy
15	Buy from Splash Maps
16	Buy from the National Trails shop
17	Buy from ViewRanger
18	Buy Now
19	Buy Online
20	e-Gift Card

Πίνακας 3: Πίνακας όρων συναλλαγής (T(payment))

Στη συνέχεια, με βάση αυτόν τον πίνακα, ελέγχεται αν στο κυρίως σώμα της σελίδας συναντάται κάποιος από τους όρους που (υπο-)δηλώνουν οικονομική συναλλαγή, και χαρακτηρίζουμε κάθε σελίδα *free* ή *not free*.

Παράλληλα, στο πλαίσιο της ίδιας διαδικασίας (**ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 1:** StructureBased Classification, **Phase 2:** *Layered Page*

Classification), κατηγοριοποιούνται σε μεγαλύτερο βάθος και οι σελίδες πλοήγησης με βάση τα αντίστοιχα βήματα. Σύμφωνα με τον προτεινόμενο αλγόριθμο, κατά το πρώτο βήμα μετρείται το πλήθος των *slashes* που χαρακτηρίζουν τον ενιαίο εντοπιστή πόρων (*URL*) της κάθε σελίδας, ώστε να μπορεί να γίνει ο διαχωρισμός τους σε *αρχικές* (*HomePages*) και *ιστοσελίδες* (*WebPages*). Το όριο που τίθεται σε αυτό το σημείο, και για αυτόν τον διαχωρισμό, είναι το *2*. Δηλαδή, όσες σελίδες έχουν στο URL τους 2 ή περισσότερα *slashes* χαρακτηρίζονται *WebPages*, διαφορετικά χαρακτηρίζονται *HomePages.* Χρειάζεται να διευκρινιστεί ότι κατά το βήμα αυτό δεν υπολογίζονται τα δύο αρχικά *slashes,* καθώς αυτά συναντώνται στα URL των περισσότερων σελίδων ως πρόθεμα ορισμένο από τους κανόνες σύνταξης τους [VII]. Έτσι, όταν ο αριθμός των *slashes* είναι ίσος ή μεγαλύτερος του 2, τότε η σελίδα χαρακτηρίζεται *WebPage* και επισημειώνεται με το πλήθος των *slashes* για να προσδιοριστεί το βάθος στο οποίο βρίσκεται μέσα στον ιστότοπο που τη φιλοξενεί. Στην αντίθετη περίπτωση, η σελίδα χαρακτηρίζεται *αρχική*, οπότε και ακολουθεί ο χαρακτηρισμός της με βάση την κατάληξή της (*suffix*), η οποία (υπο-)δηλώνει τον *τομέα* (*domain*) στο οποίο ανήκουν (π.χ. com = commercial). Για τον χαρακτηρισμό των *HomePages* με βάση την κατάληξή τους, αξιοποιείται ο πίνακας με την αναλυτική ερμηνεία της κάθε κατάληξης [II].

Συνοπτικά και επιγραμματικά, τα αποτελέσματα από την πειραματική δοκιμή του Αλγορίθμου Δομικής Κατηγοριοποίησης φαίνεται πως επιβεβαιώνουν την ορθότητα των βημάτων και των στοιχείων εκείνων πάνω στα οποία επιλέξαμε να βασιστούμε για τη λήψη αποφάσεων. Από το σύνολο των 2.330 σελίδων διαδικτύου που κλήθηκε ο αλγόριθμος που σχεδιάσαμε να κατηγοριοποιήσει με βάση τη δομή τους, κατηγοριοποιήθηκαν ως εξής*: 54% πληροφοριακές σελίδες, 16% σελίδες πλοήγησης και 18% σελίδες διάδρασης*. Κατά συνέπεια, μόλις το *12%* των σελίδων που εξετάσαμε παρέμειναν χωρίς ετικέτα κατηγορίας ως προς τον τύπο τους.

Στο σημείο αυτό ολοκληρώνεται η πειραματική εφαρμογή της δομικής κατηγοριοποίησης παρότι, στην περιγραφή του αλγορίθμου υπό μορφή ψευδοκώδικα, υπάρχει ακόμα ένα βήμα. Σκοπός αυτού του βήματος είναι η Ωδιαχείριση των περιεχόμενων σε μια σελίδα (υπερ)συνδέσμων, για αυτό και δεν

υπάρχει κάποια ενέργεια που πρέπει να γίνει στο πλαίσιο της πειραματικής εφαρμογής. Ολοκληρώνοντας τη δομική κατηγοριοποίηση των σελίδων (**ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 1:** Structure-Based Classification), προχωράμε στη θεματική (**ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 2:** Content-Based Classification).

Πρωταρχικό μέλημα για τη θεματική κατηγροιοποίηση των σελίδων διαδικτύου που απαρτίζουν το δείγμα δεδομένων, είναι να εξαχθούν και να καταγραφούν τα δύο βασικά κειμενικά στοιχεία (*Textual Elements*) για κάθε σελίδα (**Phase 1:** *Textual Elements Extraction*). Αυτά, όπως ορίζεται από την προτεινόμενη μεθοδολογία, είναι τα *P's anchorTitle* και *P's textTitle* της κάθε σελίδας. Όταν αναφερόμαστε στο *P's anchorTitle* εννοούμε τις λέξεις εκείνες που εμφανίζονται ως σύνδεσμος και μας οδηγούν στη σελίδα. Αλλιώς, από τη σκοπιά του χρήστη, είναι ο τίτλος που φαίνεται κατά την αποθήκευση της σελίδας στους σελιδοδείκτες του φυλλομετρητή. Πιο απλά, πρόκειται για τις λέξεις που εμφανίζονται ως τίτλος περιγραφής της σελίδας, και στον κώδικα html της σελίδας ορίζεται μέσα από το γνώρισμα *<a href>* από όπου μπορεί και να αντληθεί. Το στοιχείο *P's textTitle*, από την άλλη, είναι ο τίτλος που εμφανίζεται (όπου υπάρχει) στο κυρίως σώμα της σελίδας και στον hmtl κώδικα της σελίδας ορίζεται με τις ετικέτες *<h1>* έως *<h6>*, ανάλογα με την σπουδαιότητα του τίτλου [VIII]. Δηλαδή, η ετικέτα *<h1>* αφορά τον κύριο τίτλο, η ετικέτα *<h2>* τον υπότιτλο κ.ο.κ. [IX].

Ύστερα από τον εντοπισμό αυτών των στοιχείων, αποθηκεύονται ως κειμενικά στοιχεία της σελίδας, και η κάθε σελίδα επισημειώνεται με τα δικά της. Στη συνέχεια, με σκοπό την ανάθεση θέματος για κάθε σελίδα (**Phase 2:** *Theme Detection*), εντοπίζονται τα κοινά ουσιαστικά ή/και κύρια ονόματα μεταξύ *Anchor Title* και *Title*, όπως ορίζεται από την προτεινόμενη μεθοδολογία. Στις περιπτώσεις που υπάρχουν κοινά ουσιαστικά ή/και κύρια ονόματα μεταξύ αυτών των κειμενικών στοιχείων, τότε ορίζονται ως θεματικοί όροι της σελίδας. Στη συνέχεια, με βάση αυτούς τους όρους και αξιοποιώντας τις κατηγορίες της Wikipedia, προκύπτει η θεματική κατηγορία κάθε σελίδας, όπως θα δούμε στη συνέχεια.

Στις περιπτώσεις που είτε δεν είναι διαθέσιμα και τα δύο αυτά κειμενικά στοιχεία είτε δεν υπάρχουν κοινά ουσιαστικά ή/και κύρια ονόματα μεταξύ τους, γίνεται η *εξαγωγή λέξεων-κλειδιών*. Για το σκοπό αυτό, στο πλαίσιο της παρούσας πειραματικής εφαρμογής, αξιοποιείται το εργαλείο WordCounter [VI], που διατίθεται ελεύθερα στο Διαδίκτυο και το οποίο εξάγει αυτόματα τις λέξεις εκείνες που υπάρχουν στο περιεχόμενο της σελίδας σε μεγάλη συχνότητα, αλλά δεν είναι γενικώς κοινές λεκτικές μονάδες όπως είναι τα άρθρα, οι σύνδεσμοι, οι αντωνυμίες κ.ο.κ. Λαμβάνοντας υπόψη τα 5 πρώτα ουσιαστικά ή/και κύρια ονόματα των λέξεων-κλειδιών και ελέγχοντας αν υπάρχουν κοινά μεταξύ αυτών και *P's anchorTitle* ή/και *P's textTitle*, γίνεται η εξαγωγή θέματος της σελίδας, ελλείψει κάποιου εκ των προηγούμενων βημάτων.

Αν ύστερα και από το παραπάνω βήμα δεν προκύπτουν οι θεματικοί όροι, ελέγχεται αν κάποιο/α από τα ουσιαστικά ή/και κύρια ονόματα των λέξεων-κλειδιών μοιράζονται τον ίδιο ορισμό με κάποιο/α ουσιαστικό/α ή/και κύριο/α όνομα/ονόματα από αυτά του *P's anchorTitle* ή/και *P's textTitle*. Για την εκτέλεση αυτού του βήματος, αξιοποιείται η λεξικογραφική οντολογία WordNet. Όσες σελίδες παραμένουν και πάλι χωρίς θέμα, χαρακτηρίζονται *Punknown*.

Σύμφωνα με τα βήματα του προτεινόμενου αλγορίθμου, προκειμένου να ομαδοποιηθούν και να ονοματοδοτηθούν τα θέματα των σελίδων υπό επεξεργασία, αντιστοιχώνται οι κοινοί θεματικοί όροι που αναφέρονται πιο πάνω στα περιεχόμενα των κατηγοριών της Wikipedia, και η σελίδα υιοθετεί την κατηγορία στην οποία ανήκουν οι περισσότεροι από τους θεματικούς της όρους. Αν αυτοί οι όροι δεν βρίσκονται στα περιεχόμενα των κατηγοριών, επιλέγεται ο ευρύτερος από τους θεματικούς όρους της σελίδας και γίνεται αναζήτηση για άρθρο στη Wikipedia με τίτλο αυτόν τον όρο. Έτσι, σε αυτή την περίπτωση, η σελίδα υιοθετεί την κατηγορία στην οποία ανήκει το άρθρο αυτό. Για τον εντοπισμό της κατηγορίας στην οποία ανήκει το άρθρο και άρα η σελίδα, στηρίζεται στις κατηγορίες που εμφανίζονται ως ετικέτες στο κάτω μέρος του άρθρου, και οι οποίες είναι τοποθετημένες από τους επιμελητές με σειρά σχετικότητας.

Ολοκληρώνοντας τη διαδικασία της θεματικής κατηγοριοποίησης, διαπιστώνεται πως όπως αυτή ορίζεται από τον προτεινόμενο αλγόριθμο, έχει νόημα και είναι αποτελεσματική. Ενδεικτικά αναφέρουμε ότι στο σύνολο των σελίδων που επεξεργαζόμαστε, το 96,80% κατηγοριοποιήθηκαν θεματικά, ενώ μόλις το 3,20% έμειναν χωρίς θεματική κατηγορία. Αξίζει επίσης να αναφερθεί, ότι το 83,88% από τις σελίδες που κατηγοριοποιούνται θεματικά μέσα από τα βήματα του προτεινόμενου αλγορίθμου, το θέμα προκύπτει από κοινά ουσιαστικά ή/και κύρια ονόματα μεταξύ *Anchor Title* και *Text Title*, δηλαδή ήδη από το πρώτο βήμα. Από αυτό το δεδομένο, ακόμα και στην πρώτη ανάγνωση των αποτελεσμάτων της πειραματικής εφαρμογής, αποδεικνύεται πως τα στοιχεία στα οποία επιλέγουμε να βασιστούμε για τη θεματική κατηγοριοποίηση είναι αρκετά ισχυρά και έχουν νόημα. Αναλυτικότερη παρουσίαση των αποτελεσμάτων γίνεται σε επόμενη ενότητα (5.3. Πειραματικά Αποτελέσματα).

5.2.2. Επανακατηγοριοποίηση Σελίδων Διαδικτύου με βάση τον Βαθμό Αλλαγής (ALGORITHM2: ReClassification based on Change Detection)

Σε επόμενο στάδιο, γίνεται ο έλεγχος της αποτελεσματικότητας του Αλγορίθμου 2 (**Algorithm 2: Re-Classification based on Change Detection**). Για το σκοπό αυτό, εξετάζονται οι ίδιες σελίδες, ως προς τα ίδια κειμενικά στοιχεία και δομικά χαρακτηριστικά που μελετώνται κατά την εφαρμογή του Αλγορίθμου 1, και έχοντας αφήσει να μεσολαβήσει ένα τυχαίο χρονικό διάστημα. Για την εξυπηρέτηση του σκοπού αυτής της πειραματικής δοκιμής, το χρονικό αυτό διάστημα επιλέγεται να είναι ο ένας μήνας. Με αυτόν τον τρόπο, επιδιώκουμε να εντοπίσουμε τις σελίδες εκείνες που θα άλλαζαν άμεσα, πιθανότατα και συχνά.

Έτσι, έχοντας το ίδιο σύνολο πειραματικών δεδομένων, και ένα μήνα αργότερα από την εφαρμογή τού Αλγορίθμου 1, καταγράφονται οι «τιμές» για τα ίδια στοιχεία που μελετώνται και κατά την εφαρμογή του Αλγορίθμου 1. Καταγράφονται, δηλαδή, οι «τιμές» που λαμβάνινται από τη μελέτη των ίδιων χαρακτηριστικών και στοιχείων, δομικών και κειμενικών αντίστοιχα, για τις ίδιες διαδικτυακές σελίδες σε δύο διαφορετικές χρονικές στιγμές. Συγκρίνουμε τις τιμές αυτές, και όπου υπάρχει

διαφορά γίνεται η σχετική επισημείωση. Στη συνέχεια, υπολογίζεται ο βαθμός αλλαγής κάθε σελίδας, μετρώντας πόσα από το σύνολο των στοιχείων και χαρακτηριστικών που μελετώνται παρουσιάζουν αλλαγή. Καθεμιά από τις παραπάνω διαδικασίες, σύγκριση κειμενικών στοιχείων και σύγκριση δομικών χαρακτηριστικών γίνεται ξεχωριστά, όπως ορίζεται και από τον αλγόριθμο. Αυτό σημαίνει ότι υπολογίζεται ξεχωριστά ο βαθμός αλλαγής της κάθε σελίδας ως προς το θέμα της και ξεχωριστά ο βαθμός αλλαγής της ως προς τη δομή της. Σε κάθε περίπτωση, για τον υπολογισμό του βαθμού αλλαγής, μελετώνται τα αντίστοιχα στοιχεία/χαρακτηριστικά.

Προκειμένου να είναι δυνατή η σύγκριση των «τιμών» τόσο των κειμενικών στοιχείων όσο και των δομικών χαρακτηριστικών, στο πλαίσιο της πειραματικής δοκιμής της προτεινόμενης μεθοδολογίας, βασιζόμαστε στην τεχνική των *n-grams*, η οποία είναι αρκετά διαδεδομένη στο χώρο της επεξεργασίας φυσικής γλώσσας και της ανάλυσης κειμένων, όπου και αξιοποιείται ως μετρική ομοιότητας μεταξύ κειμένων. Η τεχνική αυτή ελέγχει την ομοιότητα μεταξύ ορισμένων «τεμαχίων» κειμένου, τα οποία μπορούν να είναι μεμονωμένοι χαρακτήρες (συμπεριλαμβανομένων και των κενών μεταξύ των λέξεων και των σημείων στίξης), μπορούν να είναι ζεύγη χαρακτήρων, ολόκληρες λέξεις, ολόκληρα κείμενα κ.ο.κ. Βασικά πλεονεκτήματα αυτής της τεχνικής υπολογισμού ομοιότητας είναι ότι παρέχει τη δυνατότητα εύκολων και γρήγορων υπολογισμών, είναι ανεξάρτητη από τη φυσική γλώσσα γραφής, και δεν επηρεάζεται από πιθανά ορθογραφικά λάθη.

Πολύ σημαντικό για την προτεινόμενη μεθοδολογία είναι το γεγονός ότι ανεξάρτητα από το με ποια σειρά που παρουσιάζονται οι διαδικασίες αυτές (**Procedure 1**: Re-Classification Decision based on Textual Changes και **Procedure 2:** Re-Classification Decision based on Structural Changes) στο πλαίσιο του Αλγορίθμου 2 (**Algorithm2:** ReClassification based on Change Detection), στην πραγματικότητα μπορούν να εκτελεστούν ταυτόχρονα, μιας και το αποτέλεσμα της μιας δεν επηρεάζει ούτε εξαρτάται από το αποτέλεσμα της άλλης.

Κατά την εφαρμογή του Αλγορίθμου 2, χρειάζεται να τεθεί ένα όριο σχετικά με το βαθμό αλλαγής των σελίδων. Αυτό σημαίνει ότι πρέπει να αποφασιστεί ποιο είναι

εκείνο το ποσοστό αλλαγής που θα διαχωρίσει τις σελίδες που χρειάζονται επανα-κατηγοριοποίηση, από εκείνες που δεν χρειάζονται. Ύστερα από σειρά δοκιμών, αλλά και μελέτης περιπτώσεων, και με δεδομένο το σύνολο των πειραματικών μας δεδομένων, θέτουμε σαν όριο τα 3/7 των δομικών χαρακτηριστικών και τα 3/6 των κειμενικών στοιχείων. Αυτό σημαίνει ότι αν κάποια σελίδα παρουσιάζει ποσοστό αλλαγής 50% και πάνω ως προς τα κειμενικά της στοιχεία, θεωρείται πως απαιτείται επανα-κατηγοριοποίηση ως προς το θέμα της. Ομοίως, αν παρουσιάσει ποσοστό αλλαγής 42,85% και πάνω ως προς τα δομικά της χαρακτηριστικά, απαιτείται επανα-κατηγοριοποίηση ως προς τη δομή της.

Από τα παραπάνω φαίνεται η πρακτικής σημασίας επιλογή μας κατά το σχεδιασμό της προτεινόμενης μεθοδολογίας, η θεματική και η δομική κατηγοριοποίηση να πραγματοποιούνται μέσα από δύο ξεχωριστούς και ανεξάρτητους αλγορίθμους, παρότι θα μπορούσαν να αποτελούν έναν ενιαίο αλγόριθμο.

Σε σχέση με τα ευρήματα σε αυτό το στάδιο, ενδεικτικά, αξίζει να αναφερθεί ότι μόνο το 0,8% των σελίδων υπό επεξεργασία εμφανίζει ποσοστό αλλαγής τέτοιο, τόσο σε δομικό όσο και σε θεματικό επίπεδο, ώστε να χρειάζεται να επανα-κατηγοριοποιηθεί εξ ολοκλήρου, δομικά και θεματικά. Ενώ μόνο δομικά είναι το 5% εκείνες που χρειάζονται επανα-κατηγοριοποίηση, και μόνο θεματικά το 4%.

5.2.3. Βελτιστοποίηση Επανακατηγοριοποίησης Σελίδων Διαδικτύου με βάση τον Ρυθμό Αλλαγής (Algorithm3: Optimized ReClassification based on Change's Frequency Detection)

Για την αξιολόγηση του Αλγορίθμου 3 (**Algorithm 3: Optimized Re-Classification based on Change's Frequency Detection**), στο πλαίσιο της πειραματικής δοκιμής χρειάζεται να επαναληφθεί η ίδια ακριβώς διαδικασία που γίνεται και κατά την εφαρμογή του Αλγορίθμου 2. Αυτό σημαίνει ότι λαμβάνεται ξανά το ίδιο σύνολο δεδομένων και εξετάζεται ως προς τα ίδια στοιχεία/χαρακτηριστικά, σε μεταγενέστερη χρονική στιγμή.

Πιο συγκεκριμένα, προκειμένου να είναι τυχαίες οι χρονικές αποστάσεις μεταξύ των πειραματικών δοκιμών, επιλέγεται η χρονική στιγμή που εφαρμόζεται ο Αλγόριθμος 3 να απέχει σχεδόν δύο μήνες από τη χρονική στιγμή της εφαρμογής τού Αλγορίθμου 2 (**Algorithm 2: ReClassification based on Change Detection**), και σχεδόν 3 μήνες από την εφαρμογή του Αλγορίθμου 1 (**Algorithm 1: MultiDimensional Page Classification**). Επιλέγεται, δηλαδή, να μην είναι ίσα τα χρονικά διαστήματα που μεσολαβούν μεταξύ των πειραματικών μας εφαρμογών, ώστε να αυξηθεί ο βαθμός τυχαιότητας και αντικειμενικότητας των αποτελεσμάτων. Η επιλογή αυτή δεν είναι η μοναδική, ούτε υποχρεωτική για την αποτελεσματικότητα της προτεινόμενης μεθοδολογίας. Όπως ο ορισμός των ορίων, έτσι και η επιλογή των χρονικών στιγμών εξέτασης και επανεξέτασης των σελίδων εξαρτώνται από το σύνολο των πειραματικών δεδομένων και το σκοπό της έρευνας, στο πλαίσιο της οποίας αξιοποιούνται οι προτεινόμενοι αλγόριθμοι.

Στο πλαίσιο της εφαρμογής τού Αλγορίθμου 3, χρειάζεται επίσης να τεθούν κάποια όρια. Αυτά αφορούν το διαχωρισμό μεταξύ των σελίδων που αλλάζουν καθόλου ή σπάνια (**RarelyChanging Page**), όσων αλλάζουν συχνά (**Highly-Changing Page**) και εκείνων που αλλάζουν με ρυθμό «κανονικό» και «διαχειρίσιμο» από τον προτεινόμενο αλγόριθμο (**RegularlyChanging Page**). Όσες αλλάζουν καθόλου ή σπάνια, ο προτεινόμενος αλγόριθμος τις θεωρεί στατικές, και άρα δεν τις κατηγοριοποιεί ξανά, αλλά τις αποθηκεύει σε ένα δευτερεύον ευρετήριο για περεταίρω μελέτη και έρευνα. Παράλληλα, όσες αλλάζουν συχνά, επίσης δεν τις κατηγοριοποιεί ξανά, καθώς θεωρείται σπατάλη υπολογιστικής ενέργειας και πόρων να κατηγοριοποιούνται ξανά και ξανά σελίδες οι οποίες αλλάζουν άμεσα, ενδεχομένως και πριν προλάβει να ολοκληρωθεί η διαδικασία της επανακατηγοριοποίησης. Ωστόσο, και αυτές αποθηκεύονται σε ένα δευτερεύον ευρετήριο για περεταίρω μελέτη και έρευνα. Ενώ εκείνες που αλλάζουν με ρυθμό «κανονικό» και «διαχειρίσιμο» ο αλγόριθμός μας τις στέλνει στον Αλγόριθμο 2, μέσα από τον οποίο θα υπολογιστεί ο βαθμός αλλαγής θεματικά και δομικά, και θα ακολουθηθούν τα αντίστοιχα βήματα που ορίζει ο Αλγόριθμος 2, και εφαρμόζοντας τα ίδια όρια.

Ενδεικτικά, από την πειραματική δοκιμή του Αλγορίθμου 3 και σύμφωνα με τα ευρήματα αυτής, το 41.80% των σελίδων υπό επεξεργασία παρουσιάζουν αλλαγή/ες (ως προς το θέμα τους ή/και ως προς τη δομή τους) με ρυθμό τέτοιο που χρειάζεται να σταλούν στον Αλγόριθμο 2 για επανα-κατηγοριοποίηση (**RegularlyChanging Page**).

Στις επόμενες Ενότητες, γίνεται αναλυτική παρουσίαση των αποτελεσμάτων, ενώ τα συμπεράσματα που προκύπτουν από αυτά παρουσιάζονται αναλυτικά σε επόμενο Κεφάλαιο, μαζί με την γενικότερη και συνολική αποτίμηση του έργου.

5.3. Πειραματικά Αποτελέσματα

Στο πλαίσιο αυτής της Ενότητας παρουσιάζονται τα αποτελέσματα που λαμβάνονται από την πειραματική δοκιμή της προτεινόμενης μεθοδολογίας για την αυτοματοποιημένη πολυδιάστατη κατηγοριοποίηση των σελίδων διαδικτύου, λαμβάνοντας υπόψη τον βαθμό και το ρυθμό αλλαγής των τελευταίων. Η μεθοδολογία αποτελείται από τρεις ανεξάρτητους και αυτοτελείς αλγορίθμους, και στην παρούσα Ενότητα παρουσιάζονται αναλυτικά τα αποτελέσματα από την εφαρμογή του καθενός ξεχωριστά, συνοδευόμενα από τις αντίστοιχες γραφικές τους απεικονίσεις.

Από την εφαρμογή του πρώτου αλγορίθμου (**ALGORITHM 1: Multi-Dimensional Page Classification**), αυτού της πολυδιάστατης κατηγοριοποίησης σελίδων διαδικτύου, τα αποτελέσματα που λαμβάνονται αφορούν την κατηγοριοποίηση των σελίδων ως προς τη δομή και ως προς το θέμα τους. Ξεκινώντας από τη δομική κατηγοριοποίηση (**Procedure 1:** Structure-Based Classification), οι σελίδες κατηγοριοποιούνται αρχικώς ως προς τον τύπο τους σε *πληροφοριακές*, *πλοήγησης* ή *διάδρασης* (**Phase 1:** *Page Type Recognition*). Στη συνέχεια, οι σελίδες πλοήγησης κατηγοριοποιούνται σε *αρχικές* ή *ιστοσελίδες*, και οι αρχικές κατηγοριοποιούνται με βάση τον τομέα εκείνο στον οποίο ανήκει ο ιστότοπος (**Phase 2:** *Layered Page Classification*). Παράλληλα, οι σελίδες διάδρασης κατηγοριοποιούνται με βάση τον χαρακτήρα/το σκοπό της διάδρασης με το χρήστη, καθώς και με βάση το εάν η ολοκλήρωση της διάδρασης γίνεται δωρεάν ή όχι (**Phase 2:** *Layered Page Classification*).

Δεδομένου του συνόλου δεδομένων υπό επεξεργασία κατηγοριοποιώντας τις σελίδες ως προς τον τύπο τους, το 54% των σελίδων χαρακτηρίστηκαν από τον προτεινόμενο αλγόριθμο πληροφοριακές, το 16% πλοήγησης και το 18% διάδρασης, ενώ το 12% των σελίδων έμεινε χωρίς να έχουν χαρακτηρισθεί ως προς τον τύπο τους (Πίνακας 4). Ενδεικτικά, αξίζει να αναφέρουμε πως σύμφωνα με την κατηγοριοποίηση που προέκυψε με βάση τον χαρακτηρισμό που έδωσαν οι εθελοντές σε κάθε αποθηκευμένο σελιδοδείκτη που μοιράστηκαν μαζί μας, το 58% είναι πληροφοριακές, το 23% σελίδες πλοήγησης και το 19% των σελίδων είναι σελίδες διάδρασης. Αυτή η αντιπαραβολή, μπορεί να θεωρηθεί και μια πρώτη ένδειξη αναφορικά με την ορθότητα της κατηγοριοποίησης που προέκυψε από τον αλγόριθμο που σχεδιάσαμε και προτείνουμε. Ενδελεχής και εμπεριστατωμένη αξιολόγηση των αποτελεσμάτων της πειραματικής δοκιμής τής μεθοδολογίας μας γίνεται σε επόμενη Ενότητα (5.4. Μετρικές Αξιολόγησης).

Τύπος	***Ποσοστό από την κατηγοριοποίηση του προτεινόμενου Αλγορίθμου***	***Ποσοστό από την κατηγοριοποίηση του ανθρώπου***
Informational *(Πληροφοριακές σελίδες)*	***54%***	***58%***
Navigational *(Σελίδες πλοήγησης)*	***16%***	***23%***
Transactional *(Σελίδες διάδρασης)*	***18%***	***19%***
No tag *(Σελίδες χωρίς κατηγορία)*	***12%***	***0%***
Σύνολο	***100%***	***100%***

Πίνακας 4: Αποτελέσματα δομικής κατηγοριοποίησης σελίδων ως προς τον τύπο τους (**ALGORITHM 1: MULTI-DIMENSIONAL PAGE CLASSIFICATION, Procedure 1:** Structure-Based Classification, **Phase 1:** *Page Type Recognition*)

Ύστερα από τη διαδικασία κατηγοριοποίησης των σελίδων πλοήγησης, τα αποτελέσματα είναι: 59% των σελίδων πλοήγησης του δείγματος χαρακτηρίζονται αρχικές σελίδες ιστοτόπων, ενώ το 41% ως ιστοσελίδες στο εσωτερικό κάποιου ιστοτόπου (Πίνακας 5).

Τύπος	***Ποσοστό***
HomePages *(Αρχικές σελίδες)*	***59%***
WebPages	***41%***

(Ιστοσελίδες ιστοτόπων)	
Σύνολο	***100%***

Πίνακας 5: Αποτελέσματα κατηγοριοποίησης σελίδων πλοήγησης ως προς τον τύπο τους (**ALGORITHM 1: MULTI-DIMENSIONAL PAGE CLASSIFICATION, Procedure 1:** Structure-Based Classification, **Phase 2:** *Layered Page Classification given the Type*)

Από τις σελίδες πλοήγησης, όσες χαρακτηρίζονται αρχικές από τον προτεινόμενο αλγόριθμο κατηγοριοποιούνται στη συνέχεια με βάση τον τομέα στον οποίο η κατάληξή τους (υπο-)δηλώνει πως ανήκουν. Σύμφωνα με τα αποτελέσματα ύστερα από την ολοκλήρωση και αυτής της διαδικασίας, το 63% των αρχικών σελίδων πλοήγησης του δείγματος είναι εμπορικές, το 21% αφορά σελίδες που στην κατάληξή τους δηλώνεται η χώρα στην οποία δημιουργήθηκαν και το 17% ανήκουν σε κάποιο ίδρυμα (Πίνακας 6).

Τύπος	***Ποσοστό***
Commercial *(Εμπορικές)*	***63%***
State *(Χώρα δημιουργίας)*	***21%***
Institution *(Ίδρυμα)*	***17%***
Σύνολο	***100%***

Πίνακας 6: Κατηγοριοποίηση αρχικών σελίδων πλοήγησης (**ALGORITHM 1: MULTI-DIMENSIONAL PAGE CLASSIFICATION, Procedure 1:** Structure-Based Classification, **Phase 2:** *Layered Page Classification given the Type*)

Παράλληλα, στο πλαίσιο της διαδικασίας για βαθύτερη/πληρέστερη κατηγοριοποίηση των σελίδων διάδρασης, οι ίδιες κατηγοριοποιούνται και με βάση τον σκοπό της αλληλεπίδρασης χρήστη-σελίδας. Σύμφωνα με τα αποτελέσματα που λαμβάνουμε από την πειραματική δοκιμή του προτεινόμενου αλγορίθμου, το 61% των σελίδων που χαρακτηρίζονται ως διάδρασης από τον αλγόριθμο ανήκει στην κατηγορία του ηλεκτρονικού εμπορίου, το 18% χαρακτηρίζονται σελίδες που προορίζονται για να προσφέρουν διασκέδαση, το 9% αφορά σελίδες λήψης (download) κάποιου λογισμικού ή πολυμεσικού αρχείου (multimedia file), το 7% αφορά σελίδες μέσω των οποίων πραγματοποιείται κάποια κράτηση, και το

υπόλοιπο 4% μοιράζεται ισόποσα μεταξύ σελίδων για (δια)τραπεζικές συναλλαγές και άλλα (Πίνακας 7).

Τύπος	***Ποσοστό***
Banking *(Τραπεζική συναλλαγή)*	***2%***
Booking *(Κράτηση)*	***7%***
Download *(Λήψη λογισμικού/εφαρμογών κ.α.)*	***9%***
E-commerce *(Ηλεκτρονικό εμπόριο)*	***61%***
Entertainment *(Διασκέδαση)*	***18%***
Άλλο	***2%***
Σύνολο	***100%***

Πίνακας 7: Αποτελέσματα κατηγοριοποίησης σελίδων διάδρασης με βάση τον τύπο της αλληλεπίδρασης (**ALGORITHM 1: MULTI-DIMENSIONAL PAGE CLASSIFICATION, Procedure 1:** Structure-Based Classification, **Phase 2:** *Layered Page Classification given the Type*)

Ολοκληρώνοντας τη δομική κατηγοριοποίηση των σελίδων διάδρασης, ο προτεινόμενος αλγόριθμος τις διαχωρίζει σε δωρεάν και επί πληρωμή. Στο δείγμα των πειραματικών δεδομένων, το 34% των σελίδων που χαρακτηρίζονται διάδρασης είναι δωρεάν, το 59% όχι, ενώ το 7% έχει και τα δύο είδη των υπηρεσιών (Πίνακας 8).

Τύπος	***Ποσοστό***
Free *(Δωρεάν)*	***34%***
NotFree *(Επί πληρωμή)*	***59%***
Both *(Δωρεάν και Επί πληρωμή)*	***7%***
Total	***100%***

Πίνακας 8: Κατηγοριοποίηση σελίδων διάδρασης με βάση το κόστος της ενέργειας αλληλεπίδρασης

Με βάση τα αποτελέσματα της δομικής κατηγοριοποίησης των σελίδων διαδικτύου, είναι εμφανές πως τα βήματα του προτεινόμενου αλγορίθμου, έχουν νόημα και οδηγούν στα επιδιωκόμενα αποτελέσματα. Συγκεκριμένα, η πλειονότητα των σελίδων αποδίδεται σε κάποια κατηγορία ως προς τον τύπο (συνολικά 88%), και πολύ λίγες έμειναν χωρίς τον σχετικό χαρακτηρισμό (12%). Παράλληλα, είναι πολύ

σημαντικό το γεγονός ότι το σύνολο των σελίδων που κατηγοριοποιείται δομικά σε μεγαλύτερο βάθος (σελίδες πλοήγησης και διάδρασης), μέσα από τα διαδοχικά βήματα του αλγορίθμου δομικής κατηγοριοποίησης, επίσης οδηγήθηκαν σε κάποια κατηγορία.

Μελετώντας ενδελεχώς τα αριθμητικά αποτελέσματα της πειραματικής δοκιμής της προτεινόμενης μεθοδολογίας σε σχέση με όσα έχουμε από την ομάδα των εθελοντών που χαρακτήρισαν τις σελίδες πριν τις μοιραστούν μαζί μας, κρίνεται σκόπιμο να παρουσιαστούν ορισμένα ποιοτικά συμπεράσματα που προκύπτουν από αυτά. Συγκεκριμένα, εντοπίζονται σελίδες που σύμφωνα με τον προτεινόμενο αλγόριθμο χαρακτηρίζονται πληροφοριακές, αλλά οι χρήστες τις χαρακτηρίζουν πλοήγησης και το αντίστροφο. Αυτό οφείλεται στο ότι η μπάρα πλοήγησης που είναι σταθερή στο πάνω/κάτω/δεξί/αριστερό άκρο των σελίδων, άλλοτε λαμβάνεται υπόψη από το εργαλείο που αξιοποιείται για τον υπολογισμό της αναλογίας κειμένου/συνδέσμων και άλλοτε όχι. Φαίνεται, δηλαδή, πως το εργαλείο παρουσιάζει κάποια «ασυνέπεια» στο συνυπολογισμό ή όχι της σταθερής μπάρας πλοήγησης. Στην αδυναμία αυτή οφείλεται και το λάθος στο χαρακτηρισμό μιας σελίδας πλοήγησης ως ιστοσελίδας, ενώ είναι αρχική.

Κάτι ακόμα που παρατηρείται κατά την πειραματική δοκιμή είναι πως, για ορισμένες σελίδες, το εργαλείο υπολογισμού της αναλογίας κειμένου/συνδέσμων, κατά τη δεύτερη εφαρμογή, δίνει διαφορετική αναλογία σε σχέση με τον πρώτο έλεγχο, ενώ οι επιμέρους τιμές (ποσοστό κειμένου και ποσοστό συνδέσμων) είναι ίδιες με τις προηγούμενες, και το αντίστροφο για ορισμένες άλλες. Σε αυτές τις περιπτώσεις, θεωρείται πως αυτό δεν αποτελεί πρόβλημα για την παρούσα πειραματική εφαρμογή, αφού η απόκλιση είναι τέτοια που δεν επηρεάζεται το τελικό αποτέλεσμα.

Επίσης, παρατηρείται ότι για ορισμένες σελίδες ο προτεινόμενος αλγόριθμός δεν παίρνει κάποια απόφαση για τον τύπο της σελίδας με βάση τα δομικά της χαρακτηριστικά, στις περιπτώσεις που το στοιχείο-κλειδί για αυτή την απόφαση (αναλογία κειμένου/συνδέσμων) είναι πολύ κοντά στα όρια που τίθενται στο πλαίσιο

της παρούσας πειραματικής δοκιμής. Οι χρήστες αυτές τις σελίδες τις χαρακτηρίζουν είτε πληροφοριακές είτε πλοήγησης, χωρίς να έχουν όλοι την ίδια άποψη. Αυτό σημαίνει ότι ορθώς σύμφωνα με τον προτεινόμενο αλγόριθμο όταν δεν μπορεί να ληφθεί απόφαση μεταξύ των τύπων πληροφοριακής σελίδας/σελίδας πλοήγησης, η σελίδα κατηγοριοποιείται περεταίρω δομικά με βάση και τους δύο τύπους. Από την άλλη, οι σελίδες διάδρασης φαίνεται πως εντοπίζονται σωστά από τον προτεινόμενο αλγόριθμο, ενισχύοντας τον καθοριστικό ρόλο του στοιχείου-κλειδί σε αυτό το σημείο (ο εντοπισμός τουλάχιστον ενός όρου διάδρασης που εμφανίζεται ως σύνδεσμος). Επίσης, ο λόγος για τον οποίο ορισμένες σελίδες μένουν χωρίς κατηγορία, είναι το ότι ο προτεινόμενος αλγόριθμος μας έχει αυστηρά κριτήρια για τη λήψη απόφασης.

Τα αποτελέσματα που λαμβάνονται από τη δομική κατηγοριοποίηση των σελίδων, από πλευράς ποσοστών ανά κατηγορία, δεν ταυτίζονται με αυτά που αναφέρονται στη σχετική βιβλιογραφία. Ωστόσο, κάτι τέτοιο είναι αναμενόμενο, αφού στόχος μας της παρούσας πειραματικής δοκιμής δεν είναι η κατηγοριοποίηση των διαδικτυακών δεδομένων συνολικά, αλλά ορισμένων τυχαίων σελίδων που αποτελούν τα πειραματικά δεδομένα, ώστε να ελεγχθεί η αποτελεσματικότητα της προτεινόμενης μεθοδολογίας. Επίσης, έχει σημασία το γεγονός ότι τα ποσοστά για κάθε τύπο σελίδας που αναφέρονται στη βιβλιογραφία προκύπτουν από την κατηγοριοποίηση των διαδικτυακών σελίδων με βάση τα πληροφοριακά τους αιτήματα, τον στόχο των αναζητήσεών του (search goal), ενώ το σύνολο των πειραματικών δεδομένων στο πλαίσιο της παρούσας διατριβής προκύπτει από τις αποθηκευμένες σελίδες χρηστών διαδικτύου στους σελιδοδείκτες του περιηγητή τους. Δηλαδή, από συνδέσμους που οι ίδιοι εντόπισαν ύστερα από αναζήτηση και επέλεξαν να τις αποθηκεύσουν (saved links). Παράλληλα, το διαδίκτυο αλλάζει συνεχώς, όπως και οι πληροφοριακές ανάγκες και τα πληροφοριακά αιτήματα των χρηστών. Επομένως, χρειάζεται να λάβουμε υπόψη μας πως ακόμα και τα δεδομένα της βιβλιογραφίας γύρω από αυτά, είναι δυναμικά.

Παράλληλα με την εφαρμογή του αλγορίθμου για τη δομική κατηγοριοποίηση, εφαρμόζεται και ο αλγόριθμος θεματικής κατηγοριοποίησης σελίδων διαδικτύου

(**ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 2:** Content-Based Classification). Σύμφωνα με τα πειραματικά αποτελέσματα αυτής της διαδικασίας, το 96.80% των σελίδων κατηγοριοποιήθηκε θεματικά, ενώ το 3.20% των σελίδων παρέμεινε χωρίς θεματική επισημείωση (Πίνακας 9).

Κατηγοριοποιημένες θεματικά	***Ποσοστό***
Ναι	*96.80%*
Όχι	*3.20%*
Σύνολο	***100%***

Πίνακας 9: Αποτελέσματα **κ**ατηγοριοποίησης σελίδων διαδικτύου βάσει θέματος (**ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 2:** Content-Based Classification)

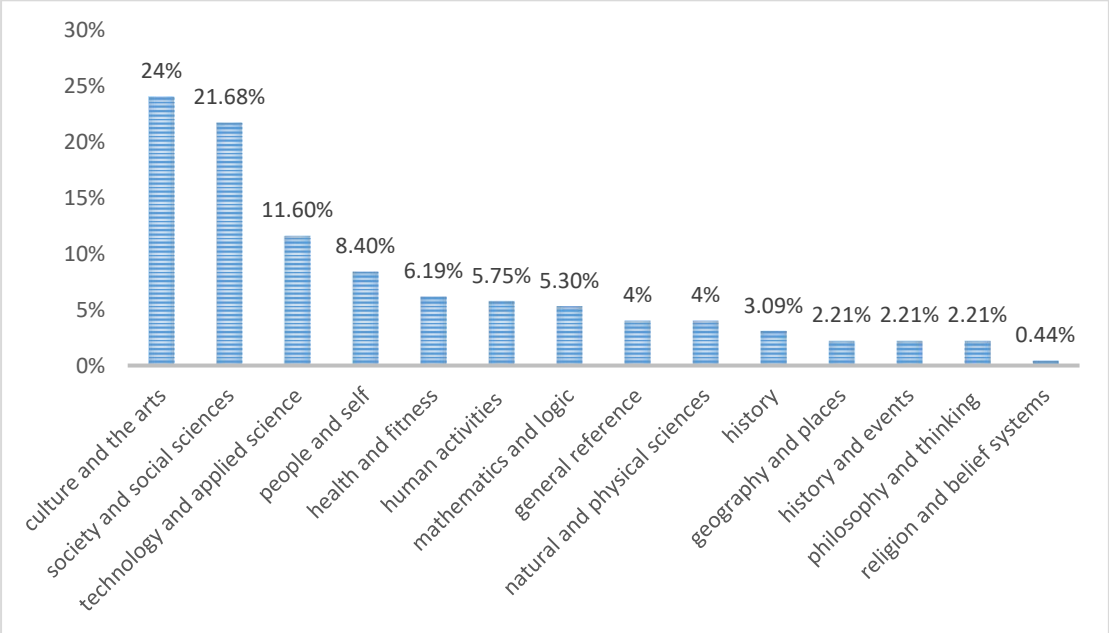
Από τα πειραματικά αποτελέσματα που λαμβάνονται από την εφαρμογή του αλγορίθμου θεματικής κατηγοριοποίησης, φαίνεται πως είναι εξαιρετικά αποτελεσματικός, καθώς σχεδόν το σύνολο των σελίδων επισημειώθηκε με κάποια θεματική κατηγορία.

Συμπληρωματικά, για την σφαιρικότερη εκτίμηση του αλγορίθμου θεματικής κατηγοριοποίησης, θεωρείται σκόπιμος ο υπολογισμός του ποσοστού των σελίδων που κατηγοριοποιούνται θεματικά από το πρώτο κιόλας βήμα της δεύτερης φάσης του αλγορίθμου, στηριζόμενος στα δύο πρώτα κατά σειρά κειμενικά στοιχεία που εξάγει ο αλγόριθμος και στα οποία καταρχήν στηρίζεται για την ανάθεση θεματικής κατηγορίας (common terms between *P's anchorTitle* and *P's textTitle*). Σύμφωνα με την παρούσα πειραματική δοκιμή, το 83.88% των σελίδων εξ όσων κατηγοριοποιούνται θεματικά, αποκτά θέμα από το πρώτο βήμα της αντίστοιχης διαδικασίας, ενώ το 16% των σελίδων που κατηγοριοποιούνται θεματικά αποκτούν θέμα κατά το δεύτερο βήμα της ίδιας διαδικασίας.

Από τα παραπάνω, τεκμηριώνεται η ορθότητα της επιλογής των κειμενικών στοιχείων που εξάγονται κατά την πρώτη φάση του αλγορίθμου προκειμένου να γίνει η θεματική κατηγοριοποίηση κατά τη δεύτερη φάση, καθώς και η σωστή η σειρά με την οποία λαμβάνονται υπόψη. Πρακτικά, αυτό σημαίνει πως έχει νόημα και είναι σωστή η σειρά με την οποία εκτελούνται τα βήματα του αλγορίθμου κατά τη διαδικασία αυτή, αφού, ήδη από το πρώτο βήμα, η πλειονότητα των σελίδων κατηγοριοποιείται

θεματικά. Η σημασία και η αποτελεσματικότητα της επιλογής των στοιχείων που λαμβάνονται υπόψη κατά τα βήματα αυτά, ενισχύεται αν λάβει κάποιος υπόψη πως είναι σχεδόν καθολικά, καθώς υπάρχουν στις περισσότερες σελίδες, και είναι ανεξάρτητα από την τεχνολογία και τη γλώσσα που χρησιμοποιούνται για την κατασκευή τους. Επομένως, η αποτελεσματικότητα της διαδικασίας είναι σε μεγάλο βαθμό ανεξάρτητη από το εκάστοτε σύνολο σελίδων υπό επεξεργασία.

Στην εικόνα που ακολουθεί, Εικόνα 10, βλέπουμε τις θεματικές κατηγορίες στις οποίες ανήκουν οι σελίδες του δείγματος δεδομένων μας, σύμφωνα με την κατηγοριοποίηση της προτεινόμενης μεθοδολογίας. Τα μεγαλύτερα ποσοστά συγκεντρώνουν οι σελίδες που αφορούν τις τέχνες και τον πολιτισμό (24%), καθώς και όσες σχετίζονται με κοινωνικές επιστήμες (21.68%). Ενώ τα μικρότερα ποσοστά οι σελίδες που σχετίζονται με τη φιλοσοφία (2.21%) και τη θρησκεία (0.44%). Υπενθυμίζουμε εδώ, ότι τα ποσοστά αυτά αφορούν αποκλειστικά το δείγμα δεδομένων που επεξεργαζόμαστε στο πλαίσιο της παρούσας πειραματικής δοκιμής. Επομένως, για κάθε άλλο σύνολο σελίδων διαδικτύου τα ποσοστά που προκύπτουν είναι διαφορετικά.



Εικόνα 10: Αποτελέσματα θεματικής κατηγοριοποίησης σελίδων, με βάση τα περιεχόμενα των κατηγοριών της Wikipedia

Σημαντική ποιοτική παρατήρηση σε αυτό το στάδιο παρουσίασης και μελέτης των πειραματικών αποτελεσμάτων, είναι ότι στο σύνολο των πειραματικών δεδομένων υπάρχουν σελίδες με διαφορετικό Url, αλλά με ίδιο θέμα, και ο αλγόριθμος το εντόπισε, δηλαδή τις ανέθεσε στην ίδια θεματική κατηγορία. Το γεγονός αυτό αποτελεί μια πρώτη ένδειξη ότι ο προτεινόμενος αλγόριθμος θεματικής κατηγοριοποίησης είναι αποτελεσματικός.

Για την αξιολόγηση του Αλγορίθμου 2 (**ALGORITHM2: Re-Classification based on Change Detection**), όπου εντοπίζονται οι σελίδες που χρειάζονται επανακατηγοριοποίηση, μελετάται ο βαθμός αλλαγής των σελίδων. Κατά την πειραματική μας εφαρμογή, εντοπίζονται αρχικώς οι σελίδες που άλλαξαν ύστερα από ένα ορισμένο χρονικό διάστημα, και υπολογίζεται το ποσοστό αλλαγής τους, στη δομή (**Procedure 2:** Re-Classification Decision based on Structural Changes) ή/και στο περιεχόμενο (**Procedure 1:** Re-Classification Decision based on Textual Changes), καθώς και ποιες εξ αυτών άλλαξαν τόσο, ώστε να χρειάζεται να επανακατηγοριοποιηθούν δομικά ή/και θεματικά.

Σύμφωνα με τα αποτελέσματα της πειραματικής δοκιμής, το ποσοστό των σελίδων για τις οποίες, σύμφωνα με τον προτεινόμενο αλγόριθμο, εντοπίζεται αλλαγή ύστερα από ένα μήνα σε σχέση με τη χρονική στιγμή που πραγματοποιείται η πρώτη κατηγοριοποίηση ως προς τη δομή τους, είναι το 56% του συνόλου των σελίδων που εξετάζονται (Πίνακας 10). Από αυτές, το 5% παρουσιάζει βαθμό αλλαγής πάνω από το όριο που τίθεται στο πλαίσιο της παρούσας πειραματικής δοκιμής, και άρα είναι αυτές που χρειάζεται να κατηγοριοποιηθούν ξανά ως προς τη δομή τους (Πίνακας 11).

Κατάσταση σελίδων	***Ποσοστό***
Αλλαγμένες 1 μήνα μετά	*56%*
Ίδιες 1μήνα μετά	*44%*
Σύνολο	***100%***

Πίνακας 10: Αποτελέσματα από τον έλεγχο δυναμικότητας δεδομένων ως προς τη δομή τους

Κατάσταση σελίδων	***Ποσοστό***
Βαθμός αλλαγής πάνω από το όριο	*5%*
Βαθμός αλλαγής κάτω από το όριο	*95%*
Σύνολο	***100%***

Πίνακας 11: Αποτελέσματα από τον υπολογισμό βαθμού αλλαγής σελίδων ως προς τη δομή τους

Μελετώντας τα αποτελέσματα από την εφαρμογή του δεύτερου αλγορίθμου σε συνδυασμό με όσα υπάρχουν από τους χρήστες, αξίζει να αναφερθεί πως σε ορισμένες σελίδες η αλλαγή που καταγράφεται στα δομικά χαρακτηριστικά οφείλεται στο ότι η σελίδα δεν υπάρχει πια. Επίσης, ορισμένες σελίδες, ενώ κατά την πρώτη κατηγοριοποίησή τους χαρακτηρίζονται από τον αλγόριθμο ως σελίδες διάδρασης, στη δεύτερη χαρακτηρίζονται πληροφοριακές. Αυτό συμβαίνει επειδή το προϊόν δεν είναι πια διαθέσιμο, άρα δεν μπορεί να γίνει αγορά, και πλέον αναφέρονται μόνο οι πληροφορίες του προϊόντος.

Άλλη σημαντική παρατήρηση από τη μελέτη των αλλαγών στη δομή των σελίδων, είναι το γεγονός ότι στις περιπτώσεις που μια σελίδα αλλάζει μόνο ως προς την αναλογία κειμένου/συνδέσμων, το ποσοστό αλλαγής της σελίδας είναι 1/7 (14.28%). Αυτό σημαίνει ότι ο αλγόριθμος δεν θα κρίνει απαραίτητη την επανα-κατηγοριοποίηση της σελίδας ως προς τη δομή, τη στιγμή που αυτή η αλλαγή μπορεί να είναι καθοριστική για τον τύπο της σελίδας. Δηλαδή, μπορεί η αναλογία κειμένου/συνδέσμων να έχει αλλάξει τόσο που να επηρεάζει την απόφαση για τον τύπο της σελίδας, αλλά ο προτεινόμενος αλγόριθμος να μην την επανα-κατηγοριοποιήσει, επειδή το ποσοστό αλλαγής ως προς τη δομή της είναι πολύ μικρό. Η αδυναμία αυτή του προτεινόμενου αλγορίθμου, μπορεί να αντιμετωπιστεί με την τεχνική των *weighted elements*. Μπορούμε, δηλαδή, να χαρακτηρίσουμε κάθε στοιχείο που λαμβάνεται υπόψη με ένα «βάρος», ώστε το ποσοστό αλλαγής των σελίδων να προκύπτει συνυπολογίζοντας το βάρος κάθε στοιχείου που αλλάζει. Με τον τρόπο αυτό, αυξάνεται η αποτελεσματικότητα του αλγορίθμου και εξοικονομούνται υπολογιστικοί πόροι και δύναμη. Αναλόγως μπορούμε να διαχειριστούμε τις αλλαγές στους όρους διάδρασης. Δηλαδή, θα μπορούσε άλλη βαρύτητα να έχει ο εντοπισμός αλλαγής σε μια σελίδα διάδρασης, όταν αυτή αφορά την απουσία τους ενώ πριν υπήρχαν, και το αντίστροφο, και άλλη βαρύτητα η αλλαγή των όρων ως προς το περιεχόμενό τους.

Παράλληλα, στο πλαίσιο της πειραματικής δοκιμής, μελετάται η δυναμικότητα των ίδιων σελίδων ως προς το θέμα τους. Από αυτή τη διαδικασία, προκύπτει πως το 10% των σελίδων παρουσιάζει αλλαγή στα κειμενικά τους στοιχεία (Πίνακας 12), και από

αυτές το 4% αλλάζει τόσο, που χρειάζεται να κατηγοριοποιηθεί ξανά θεματικά (Πίνακας 13).

Κατάσταση σελίδων	***Ποσοστό***
Αλλαγμένες 1 μήνα μετά	*10%*
Ίδιες 1 μήνα μετά	*90%*
Σύνολο	***100%***

Πίνακας 12: Αποτελέσματα από τον έλεγχο δυναμικότητας δεδομένων ως προς το περιεχόμενό τους

Κατάσταση σελίδων	***Ποσοστό***
Βαθμός αλλαγής πάνω από το όριο	*4%*
Βαθμός αλλαγής κάτω από το όριο	*96%*
Σύνολο	***100%***

Πίνακας 13: Αποτελέσματα από τον έλεγχο υπολογισμού βαθμού αλλαγής σελίδων ως προς το περιεχόμενό τους

Από τα παραπάνω, η πρώτη διαπίστωση που γίνεται είναι πως (ποσοστιαία) η πλειονότητα των σελίδων που αλλάζουν ως προς τη δομικά τους χαρακτηριστικά δεν χρειάζονται επανακατηγοριοποίηση, ενώ η πλειονότητα εξ όσων αλλάζουν ως προς τα θεματικά τους στοιχεία χρειάζεται να επανακατηγοριοποιηθούν. Αυτό φαίνεται να εξηγείται από το γεγονός ότι μεγάλου βαθμού αλλαγές στη δομή μιας σελίδας, σηματοδοτούν συχνά αλλαγή στον δομικό τους χαρακτηρισμό, ο οποίος συνδέεται με τον σκοπό της δημιουργίας τους και το σκοπό της ύπαρξής τους.

Τέλος, κατά την ολοκλήρωση της πειραματικής δοκιμής της προτεινόμενης μεθοδολογίας, παρατηρείται η «συμπεριφορά» των ίδιων σελίδων σε τυχαίες χρονικές στιγμές, μελετώντας τα ίδια δομικά και κειμενικά τους στοιχεία. Εφαρμόζοντας τον Αλγόριθμο 3 (**ALGORITHM 3: Optimized Re-Classification based on Change's Frequency Detection**) επιδιώκεται η βελτιστοποίηση της λειτουργίας τού Αλγορίθμου 2 (**ALGORITHM2: Re-Classification based on Change Detection**), καθώς μέσα από τον πρώτο (Αλγόριθμος 3) εντοπίζεται και υπολογίζεται ο ρυθμός μεταβολής των σελίδων, με αποτέλεσμα να «επιλέγονται» εκείνες μόνο που «χρειάζεται» να «σταλούν» στον δεύτερο (Αλγόριθμος 2). Με τον τρόπο αυτό, εξασφαλίζεται η ορθή χρήση υπολογιστικών πόρων.

Σύμφωνα με τα αποτελέσματα της πειραματικής δοκιμής, το 36.80% του συνόλου των πειραματικών μας δεδομένων χαρακτηρίζεται από τον προτεινόμενο αλγόριθμο

ως υψηλής συχνότητας μεταβαλλόμενες σελίδες (*HighlyChanging Page*) ως προς τη δομή τους, το 32.80% ως σπανίως μεταβαλλόμενες σελίδες (*RarelyChanging Page*), και το 30.40% ως μέτριας συχνότητας μεταβαλλόμενες σελίδες (*RegularlyChanging Page*) (Πίνακας 14). Αυτό σημαίνει πως για το 30.40% των σελίδων υπάρχει λόγος να ελεγχθεί ο βαθμός αλλαγής τους μέσα από τον Αλγόριθμο 2, και άρα χωρίς αυτό το διαχωρισμό, θα γίνει ο σχετικός έλεγχος και στο υπόλοιπο 69.60% χωρίς να υπάρχει ουσιαστικός λόγος, αφού οι τελευταίες είτε παραμένουν στατικές με την πάροδο του χρόνου, είτε αλλάζουν τόσο συχνά που είναι «μάταιο» για τον αλγόριθμό μας να κατηγοριοποιούνται ξανά και ξανά.

Συχνότητα αλλαγής	***Ποσοστό***
HighlyChanging Pages *(Σελίδες που αλλάζουν συχνά)*	*36.80%*
RarelyChanging Pages *(Σελίδες που αλλάζουν σπάνια)*	*32.40%*
RegularlyChanging Pages *(Σελίδες που αλλάζουν με μέτρια συχνότητα)*	*30.80%*
Σύνολο	***100%***

Πίνακας 14: Αποτελέσματα από τον υπολογισμό του ρυθμού αλλαγής δομικών χαρακτηριστικών των σελίδων

Σχετικά με το περιεχόμενο των σελίδων, σύμφωνα με τα αποτελέσματα της πειραματικής εφαρμογής, το 1% του συνόλου των πειραματικών μας δεδομένων χαρακτηρίζεται από τον αλγόριθμό μας ως υψηλής συχνότητας μεταβαλλόμενες σελίδες (*HighlyChangingPages*) ως προς το περιεχόμενό τους, το 87% ως σπανίως μεταβαλλόμενες σελίδες (*RarelyChanging Pages*), και το 12% ως μέτριας συχνότητας μεταβαλλόμενες σελίδες (*RegularlyChanging Pages*) (Πίνακας 15). Αυτό σημαίνει πως για το 12% των σελίδων υπάρχει λόγος να ελεγχθεί ο βαθμός αλλαγής τους μέσα από τον Αλγόριθμο 2, και άρα χωρίς αυτό το διαχωρισμό, θα γινόταν ο σχετικός έλεγχος και στο υπόλοιπο 88% χωρίς να υπάρχει ουσιαστικός λόγος, αφού οι τελευταίες είτε παραμένουν στατικές με την πάροδο του χρόνου, είτε αλλάζουν τόσο συχνά που είναι «μάταιο» για τον αλγόριθμό μας να κατηγοριοποιούνται ξανά και ξανά.

Συχνότητα αλλαγής	***Ποσοστό***
HighlyChanging Pages *(Σελίδες που αλλάζουν συχνά)*	*1%*
RarelyChanging Pages *(Σελίδες που αλλάζουν σπάνια)*	*87%*
RegularlyChanging Pages	*12%*

(Σελίδες που αλλάζουν με «ελεγχόμενη» συχνότητα)	
Σύνολο	***100%***

Πίνακας 15: Αποτελέσματα από τον υπολογισμό του ρυθμού αλλαγής στο περιεχόμενο των σελίδων

Μελετώντας τα αποτελέσματα της πειραματικής εφαρμογής που παρουσιάζονται στην παρούσα ενότητα, διαπιστώνονται η χρησιμότητα και η αποτελεσματικότητα της προτεινόμενης μεθοδολογίας στην πολυδιάστατη κατηγοριοποίηση σελίδων διαδικτύου. Ιδιαίτερης σημασίας κρίνεται το γεγονός ότι καταφέρνει να αντιμετωπίσει και τη δυναμικότητα των σελίδων, αφού υπολογίζει το βαθμό αλλαγής, αλλά και την άτακτη χρονικά εμφάνιση αυτών των αλλαγών, αφού λαμβάνει υπόψη και τον ρυθμό αλλαγής.

Στην επόμενη ενότητα, αξιολογούμε ποιοτικά την αποτελεσματικότητα (effectiveness) της προτεινόμενης μεθοδολογίας με βάση τις μετρικές ανάκλησης (recall) και ακρίβειας (precision) που αξιοποιούνται ευρέως για αυτό το σκοπό στο χώρο της κατηγοριοποίησης δεδομένων.

5.4. Μετρικές Αξιολόγησης

Ολοκληρώνοντας την πειραματική εφαρμογή της προτεινόμενης μεθοδολογίας, και έχοντας στη διάθεσή μας τα αποτελέσματά της, πραγματοποιούμε έλεγχο της πληρότητας και της ορθότητας των αποτελεσμάτων που λαμβάνουμε από αυτήν στο πλαίσιο της πειραματικής δοκιμής, βασιζόμενοι στις μετρικές της *ανάκλησης* (recall) και της *ακρίβειας* (precision), ευρέως γνωστές στο χώρο της ανάκτησης πληροφορίας και των συστημάτων μηχανικής μάθησης για την κατηγοριοποίηση των δεδομένων [9]. Σκοπός της εφαρμογής αυτών των μετρικών είναι ο ποιοτικός έλεγχος των αποτελεσμάτων που δίνονται από την εφαρμογή ενός συστήματος κατηγοριοποίησης.

Με τον τύπο της Ανάκλησης, Τύπος 1, υπολογίζουμε τον λόγο των σελίδων που ο αλγόριθμός μας κατηγοριοποίησε σωστά (TP: true positives), προς το σύνολο των σελίδων που εξετάστηκαν (TP: true positives + FN: false negatives). Έτσι, ελέγχουμε

την ικανότητα του αλγορίθμου να εντοπίζει μία κατηγορία για κάθε σελίδα που επεξεργάζεται.

$$\boldsymbol{Recall = \frac{True\ Positive}{True\ Positives + False\ Negatives}}$$

Τύπος 1: Τύπος υπολογισμού ανάκλησης αποτελεσμάτων

Με τον τύπο της Ακρίβειας, **Τύπος 2**, υπολογίζουμε τον λόγο των σελίδων που ο αλγόριθμός μας κατηγοριοποίησε σωστά (TP: true positives), προς το σύνολο των σελίδων που κατηγοριοποιήθηκαν από τον αλγόριθμο (TP: true positives + FN: false positives). Με τον τρόπο αυτό, ελέγχουμε την ικανότητα του αλγορίθμου να εντοπίζει τη σωστή κατηγορία για κάθε σελίδα που επεξεργάζεται.

$$\boldsymbol{Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}}$$

Τύπος 2: Τύπος υπολογισμού ακρίβειας αποτελεσμάτων

Έχοντας συλλέξει τα πειραματικά δεδομένα με τις σχετικές επισημειώσεις των εθελοντών που μας τα παρέχουν, και έχοντας ορίσει τις τιμές των ορίων που χρειάζεται ο προτεινόμενος αλγόριθμος για να πάρει αποφάσεις, πραγματοποιείται η πειραματική δοκιμή, όπως αυτή περιγράφεται στην προηγούμενη Ενότητα (5.2. Περιγραφή Πειραματικής Δοκιμής). Στη συνέχεια, συγκρίνουμε τις δομικές και τις θεματικές κατηγορίες που ο προτεινόμενος αλγόριθμος αποδίδει σε κάθε σελίδα με αυτές που μας είχαν δώσει οι εθελοντές.

Πιο συγκεκριμένα, για την ποιοτική αξιολόγηση των αποτελεσμάτων της δομικής κατηγοριοποίησης που προκύπτει από τον αλγόριθμο, συγκρίνουμε τον δομικό τύπο που μας είχαν εξαρχής δώσει οι εθελοντές για κάθε σελίδα που μοιράστηκαν μαζί μας, με τους αντίστοιχους που παίρνουμε από τον προτεινόμενο αλγόριθμο μετά την εφαρμογή του στις ίδιες σελίδες. Όπου ο δομικός χαρακτηρισμός που λαμβάνεται από τον αλγόριθμο ταυτίζεται με εκείνον των εθελοντών, θεωρείται ως αληθές θετικό (*true positive*) αποτέλεσμα για την κατηγορία, εναλλακτικά όταν σε μία σελίδα αποδίδεται λάθος κατηγορία, θεωρείται ψευδώς θετικό (*false positive*) για την

κατηγορία που αποδόθηκε και ψευδώς αρνητικό (*false negative*) για την κατηγορία στην οποία ανήκει η σελίδα πραγματικά σύμφωνα με τη γνώμη των χρηστών.

Στον πίνακα που ακολουθεί, **Πίνακας 16**, φαίνονται αναλυτικά τα αποτελέσματα της αξιολόγησης της δομικής κατηγοριοποίησης (**ALGORITHM 1: Multi-Dimensional Page Classification**, **Procedure 1:** Structure-based Classification). Μελετώντας τα στοιχεία, φαίνεται ότι η απόδοση της προτεινόμενης μεθοδολογίας είναι πολύ καλή. Συγκεκριμένα, κατά τη διαδικασία κατηγοριοποίησης των σελίδων με βάση τον τύπο τους (**ALGORITHM 1: Multi-Dimensional Page Classification**, **Procedure 1:** Structure-based Classification, **Phase 1**: *Page Type Recognition*), βλέπουμε ότι ο αλγόριθμος είναι αποδοτικότερος κατά τον εντοπισμό των σελίδων διάδρασης, γεγονός που αποδεικνύει την ορθότητα και την ισχύ του στοιχείου (σ.σ. όρος/όροι διάδρασης στο περιεχόμενο της σελίδας) βάσει του οποίου γίνεται ο εντοπισμός τους. Παράλληλα, από την αξιολόγηση των αποτελεσμάτων της βαθύτερης κατηγοριοποίησης (**ALGORITHM 1: Multi-Dimensional Page Classification, Procedure 1:** Structure-Based Classification, **Phase 2:** *Layered Page Classification*), παρατηρείται ότι η απόδοση του αλγορίθμου είναι καλύτερη για τις σελίδες διάδρασης από ό,τι για τις σελίδες πλοήγησης, γεγονός που εξηγείται από την προηγούμενη διαπίστωση. Δηλαδή, η όποιες αδυναμίες στη βαθύτερη κατηγοριοποίηση των σελίδων πλοήγησης οφείλονται στις αδυναμίες των αποτελεσμάτων του αλγορίθμου από την προηγούμενη φάση, όπου γίνεται ο διαχωρισμός μεταξύ πληροφοριακών και πλοήγησης, και όχι στα δομικά χαρακτηριστικά που επιλέγουμε κατά το σχεδιασμό της μεθοδολογίας για τη βαθύτερη κατηγοριοποίηση των σελίδων πλοήγησης.

	Precision *(Ανάκληση)*	***Recall*** *(Ακρίβεια)*
Informational	*0.96*	*0.87*
Navigational	*0.97*	*0.65*
Transactional	*0.93*	*1*
Average	***0.95333***	***0.84***
Navigational Pages		
HomePages	*0.91*	***0.53***
WebPages	*1*	***0.8***
Average	***0.955***	***0.665***
HomePages		
Commercial	*1*	*0.58*
State	*1*	*0.57*

Institution	*1*	*0.44*
Average	***1***	***0.53***
TransactionalPages		
Banking	*1*	*1*
Booking	*1*	*1*
Download	*1*	*1*
E-commerce	*0.96*	*1*
Entertainment	*0.96*	*1*
Other	*1*	*1*
Average	***0.98667***	***1***
TransactionCost		
Free	*1*	*1*
NotFree	*1*	*1*
Both	*1*	*1*
Average	***1***	***1***

Πίνακας 16: Ανάκληση και Ακρίβεια δομικής κατηγοριοποίησης

Ανάλογη διαδικασία ακολουθείται για την ποιοτική αξιολόγηση των αποτελεσμάτων και της θεματικής κατηγοριοποίησης (**ALGORITHM 1: Multi-Dimensional Page Classification**, **Procedure 2:** Content-Based Classification). Με άλλα λόγια, ελέγχεται η ικανότητα του αλγορίθμου να εντοπίζει το θέμα της κάθε σελίδας. Για το σκοπό αυτό, συγκρίνεται η θεματική κατηγορία που δίνουν οι εθελοντές για κάθε σελίδα με εκείνη που εξάγει ο αλγόριθμος κατά την πειραματική δοκιμή του. Όπου αυτές ταυτίζονται, το θέμα θεωρείται αληθώς θετικό (true positive), εναλλακτικά όταν σε μία σελίδα αποδίδεται λάθος κατηγορία, θεωρείται ψευδώς θετικό (false positive) για την κατηγορία που αποδόθηκε και ψευδώς αρνητικό (false negative) για την κατηγορία στην οποία ανήκει η σελίδα πραγματικά σύμφωνα με τη γνώμη των χρηστών. Μάλιστα, δεδομένου ότι για την ονοματοδοσία του θέματος των σελίδων στηριζόμαστε στις κατηγορίες της εγκυκλοπαίδειας Wikipedia, η διαδικασία αυτή γίνεται εξαιρετικά απλή.

Ύστερα από την αξιολόγηση των αποτελεσμάτων της θεματικής κατηγοριοποίησης, συγκρίνοντάς τα με τα θέματα που δίνονται από τους εθελοντές, το 97% των σελίδων που χαρακτηρίζονται θεματικά από τον αλγόριθμο, κατηγοριοποιήθηκε σωστά και το 3% λανθασμένα (Πίνακας 17). Έτσι, φαίνεται πως η προτεινόμενη μεθοδολογία είναι αποδοτική και κατά την θεματική κατηγοριοποίηση.

TruePositive	*97%*

(Αληθώς θετικά)	
FalsePositive	
(Ψευδώς θετικά)	*3%*

Πίνακας 17: Αξιολόγηση αποτελεσμάτων θεματικής κατηγοριοποίησης

Επομένως, συνολικά η μεθοδολογία που προτείνεται στο πλαίσιο της παρούσας διατριβής για την πολυδιάστατη κατηγοριοποίηση των σελίδων διαδικτύου, κρίνεται ικανοποιητικά αποδοτική.

5.5. Συγκριτική Μελέτη

Σε προηγούμενη Ενότητα (5.2. Περιγραφή Πειραματικής Δοκιμής) περιγράφεται η πειραματική δοκιμή που πραγματοποιείται για τον έλεγχο της αποτελεσματικότητας της μεθοδολογίας που σχεδιάζεται και προτείνεται στο πλαίσιο της παρούσας διατριβής. Συμπληρωματικά σε αυτή τη διαδικασία, εκτελείται ακόμα μία, κατά την οποία κατηγοριοποιείται το ίδιο σύνολο σελίδων σύμφωνα με έναν από τους βασικούς αλγόριθμους κατηγοριοποίησης. Με αυτόν τον τρόπο, δίνεται η δυνατότητα για την πραγματοποίηση ενός συγκριτικού ελέγχου της αποτελεσματικότητας της προτεινόμενης μεθοδολογίας. Συγκεκριμένα, στο πλαίσιο αυτής της διαδικασίας, υιοθετούμε έναν *k*-NN αλγόριθμο, ο οποίος είναι ο πιο αναγνωρισμένος και ευρέως χρησιμοποιούμενος αλγόριθμος στο ερευνητικό πεδίο της κατηγοριοποίησης δεδομένων.

Παραδοσιακά, ένας *k*-NN αλγόριθμος στηρίζεται στην αρχή των πλησιέστερων γειτόνων. Εν συντομία, σύμφωνα με τη λειτουργία ενός *k*-NN αλγορίθμου, αρχικά επιλέγονται οι *k* πλησιέστεροι γείτονες που αποτελούν το δείγμα εκπαίδευσης, και στη συνέχεια προβλέπει την κατηγορία εκείνη από το δείγμα εφαρμογής με τη μεγαλύτερη τάξη μεταξύ των πλησιέστερων κατηγοριών από το δείγμα εκπαίδευσης. Ο *k*-NN αλγόριθμος υπολογίζει την απόσταση μεταξύ κάθε δείγματος εκπαίδευσης και δείγματος εφαρμογής, και επιστρέφει τα *k* πλησιέστερα δείγματα [86].

Για την εφαρμογή του *k*-NN αλγορίθμου, χρειάζεται να επιλεγεί η κατάλληλη τιμή του παράγοντα *k*, καθώς ο βαθμός επιτυχίας της κατηγοριοποίησης εξαρτάται σε μεγάλο βαθμό από αυτή την τιμή. Υπάρχουν αρκετοί τρόποι που μπορούν να

χρησιμοποιηθούν για οριστεί η τιμή του *k*, με απλούστερο όλων εκείνον όπου ο αλγόριθμος δοκιμάζεται αρκετές φορές με διαφορετικές τιμές *k*, καταλήγοντας έτσι σε εκείνη που είναι πιο κατάλληλη. Παρόλο που ο *k*-NN αλγόριθμος έχει το μειονέκτημα ότι η απόδοσή του είναι χαμηλή, χρησιμοποιείται ευρέως στην κατηγοριοποίηση κειμένων, χάρη στην απόλυτη εξάρτησή του από το δείγμα εκπαίδευσης και το δείγμα εφαρμογής [33].

Επιλέγεται η σύγκριση της αποτελεσματικότητας του προτεινόμενου αλγορίθμου με αυτή του *k*-NN αλγορίθμου, καθώς παρουσιάζουν κοινά βασικά χαρακτηριστικά. Ο αλγόριθμος κατηγοριοποίησης που σχεδιάζουμε και προτείνουμε στο πλαίσιο αυτής της διατριβής, είναι γραμμικός, όπως και ο *k*-NN αλγόριθμος. Επίσης, είναι και οι δύο απλοί στην εφαρμογή τους και αποδίδουν μία κατηγορία σε κάθε αντικείμενο-δεδομένο προς κατηγοριοποίηση. Παρόλα αυτά, έχουν και μία βασική διαφορά. Ο *k*-NN αλγόριθμος είναι ένας αλγόριθμος μηχανικής μάθησης, και άρα απαιτεί φάση «εκπαίδευσης», ενώ ο προτεινόμενος σε αυτή τη διατριβή λειτουργεί χωρίς έχει προηγηθεί φάση εκπαίδευσης. Προκειμένου να εξαλειφθεί ο αντίκτυπος αυτής της διαφοράς, δίνουμε στο *k* τη μικρότερη δυνατή τιμή (*k*=3) μεταξύ αυτών που βάσει βιβλιογραφίας είναι οι ιδανικές [27] [89]. Με αυτόν τον τρόπο, δημιουργούνται παρόμοιες συνθήκες λειτουργίας, κάνοντας και τον *k*-NN αλγόριθμο να λειτουργεί σε πραγματικό χρόνο. Με άλλα λόγια, αφήνεται ο *k*-NN αλγόριθμος να λειτουργήσει με την μικρότερου δυνατού βαθμού εκπαίδευση, προκειμένου τα αποτελέσματά του να είναι συγκρίσιμα με αυτά του προτεινόμενου αλγορίθμου. Συμπληρωματικά, τα πειραματικά δεδομένα (σελίδες προς κατηγοριοποίηση) είναι τα ίδια και στις δύο περιπτώσεις. Μιλώντας πιο συγκεκριμένα για την περίπτωση του *k*-NN αλγορίθμου, ο αλγόριθμος εκπαιδεύεται με βάση τις θεματικές κατηγορίες των 3 πιο αντιπροσωπευτικών σελίδων διαδικτύου κάθε κατηγορίας.

Τέλος, χρειάζεται να διευκρινιστεί πως η παρούσα συγκριτική μελέτη αφορά μόνο τη θεματική κατηγοριοποίηση, αφού αυτό είναι και το μόνο κομμάτι της προτεινόμενης μεθοδολογίας που τα αποτελέσματά του μπορούν να συγκριθούν με αυτά μιας άλλης. Στην επόμενη Ενότητα, 5.6. Αποτελέσματα συγκριτικής πειραματικής μελέτης, παρουσιάζονται αναλυτικά και συγκριτικά τα αποτελέσματα της θεματικής

κατηγοριοποίσησης των σελίδων που αποτελούν το σύνολο των πειραματικών δεδομένων, ακολουθώντας τα βήματα του προτεινόμενου αλγόριθμου και αξιοποιώντας τον *k*-nn αλγόριθμο.

5.6. Αποτελέσματα συγκριτικής πειραματικής μελέτης

Όπως φαίνεται αναλυτικά στις Ενότητες 5.3. Πειραματικά Αποτελέσματα και 5.4. Μετρικές Αξιολόγησης, ο προτεινόμενος αλγόριθμος καταφέρνει να κατηγοριοποιήσει το 94% των σελίδων υπό επεξεργασία, εκ των οποίων το 95% κατηγοριοποιούνται στη σωστή κατηγορία. Την ίδια στιγμή, ο αλγόριθμος *k*-NN καταφέρνει να κατηγοριοποιήσει το 80% των ίδιων σελίδων υπό επεξεργασία, εκ των οποίων το 97% κατηγοριοποιούνται στη σωστή κατηγορία. Εκτιμώντας συνολικά τα αποτελέσματα της συγκριτικής μελέτης, φαίνεται πως ο προτεινόμενος αλγόριθμος έχει ανάλογη απόδοση με έναν τυπικό αλγόριθμο κατηγοριοποίησης. Στους πίνακες που ακολουθούν, Πίνακας 18 και Πίνακας 19, παρουσιάζονται αναλυτικά τα αποτελέσματα της της θεματικής κατηγοριοποίησης του προτεινόμενου αλγορίθμου συγκριτικά με αυτά του *k*-NN.

Κατηγορίες	**Ποσοστό σελίδων ανά κατηγορία σύμφωνα με τον αλγόριθμο**	**Ποσοστό σελίδων ανά κατηγορία σύμφωνα με τον αλγόριθμο *k*-NN**
culture and the arts	*25.20%*	*9.00%*
society and social sciences	*20.00%*	*23.00%*
people and self	*12.00%*	*16.20%*
technology and applied science	*10.00%*	*11.00%*
health and fitness	*7.30%*	*8.50%*
human activities	*5.30%*	*8.50%*
mathematics and logic	*4.70%*	*5.40%*
general reference	*4.60%*	*7.90%*
natural and physical sciences	*4.10%*	*4.90%*
geography and places	*2.20%*	*1.80%*
history and events	*2.20%*	*1.80%*
philosophy and thinking	*2.00%*	*1.80%*
religion and belief systems	*0.40%*	*0.20%*

Πίνακας 18: Συγκριτική παρουσίαση αποτελεσμάτων θεματικής κατηγοριοποίησης

Κατηγορίες	**Αληθώς θετικά (true positive) (αλγόριθμος)**	**Αληθώς θετικά (true positive) (*k*-NN)**
culture and the arts	*100%*	*100%*
society and social sciences	*95.74%*	*97.82%*
people and self	*96.15%*	*96.80%*
technology and applied science	*95.65%*	*100%*
health and fitness	*100%*	*100%*
human activities	*85.71%*	*94.11%*
mathematics and logic	*100%*	*90.90%*
general reference	*100%*	*87.50%*
natural and physical sciences	*88.88%*	*100%*
geography and places	*80%*	*100%*
history and events	*80%*	*100%*
philosophy and thinking	*100%*	*100%*
religion and belief systems	*100%*	*50%*
Average	***94%***	***94%***

Πίνακας 19: Συγκριτική παρουσίαση αξιολόγησης αποτελεσμάτων θεματικής κατηγοριοποίησης

Μελετώντας τον Πίνακας 18, παρατηρείται ότι το ποσοστό των σελίδων που αποδίδονται σε κάθε κατηγορία από τον προτεινόμενο αλγόριθμο και αυτό των σελίδων που αποδίδονται από τον *k*-NN είναι πολύ κοντά μεταξύ τους, στις περισσότερες των περιπτώσεων. Ωστόσο, παρατηρείται μια αξιοσημείωτη απόκλιση στην κατηγορία *culture and arts*, όπου ο *k*-NN αποδίδει αισθητά λιγότερες. Αυτό είναι κάτι αναμενόμενο, αφού πρόκειται για την κατηγορία που συγκεντρώνει τις περισσότερες σελίδες από το δείγμα δεδομένων ενώ η τιμή του *k* είναι η μικρότερη αποδεκτή, όπως έχει ήδη εξηγηθεί νωρίτερα.

Παρατηρώντας τον Πίνακας 19, γίνεται φανερό πως, κατά μέσο όρο, η συνολική απόδοση των δύο αλγορίθμων κατηγοριοποίησης είναι ίδια παρά τις επιμέρους διαφοροποιήσεις. Το γεγονός αυτό είναι αξιοσημείωτο, αφού είναι και αυτό που δίνει επιπλέον αξία στην προτεινόμενη μεθοδολογία. Αυτό που κάνει όμως τον αλγόριθμό μας να ξεχωρίζει ακόμα και σε αυτή την περίπτωση, είναι το γεγονός ότι δεν χρειάζεται εκπαίδευση και καταφέρνει να αντιμετωπίσει τη δυναμική φύση των δεδομένων. Τα δύο αυτά χαρακτηριστικά είναι σημαντικά πλεονεκτήματα, αφού χάρη στο πρώτο εξοικονομείται χρόνος και χάρη στο δεύτερο η κατηγοριοποίηση των δεδομένων επικαιροποιείται όπου και όταν χρειάζεται. Η εξοικονόμηση χρόνου

οφείλεται στο γεγονός ότι δεν απαιτείται φάση εκπαίδευσης, άρα ούτε συγκέντρωση και επεξεργασία δεδομένων εκπαίδευσης, και τα αποτελέσματα της κατηγοριοποίησης δίνονται απευθείας, χωρίς να χρειάζονται επιπλέον πόροι.

Σε κάθε περίπτωση, χρειάζεται να τονιστεί πως η παραπάνω εκτίμηση δεν έχει καθολική ισχύ ούτε είναι ανεξάρτητη από παραμέτρους. Κι αυτό γιατί η απόδοση της προτεινόμενης μεθοδολογίας συνδέεται άρρηκτα με το σύνολο δεδομένων εφαρμογής και δεν μπορεί να γενικευτεί ως απόλυτο γεγονός. Παραδοσιακά, αυτό που είναι γνωστό, είναι πως ένας κατηγοριοποιητής μηχανικής μάθησης παρουσιάζει μια πιο σταθερή συμπεριφορά και τα αποτελέσματά του μπορούν να προβλεφθούν, ιδίως όταν του παρέχεται ένα πολύ καλό δείγμα εκπαίδευσης.

ΚΕΦΑΛΑΙΟ 6: Συμπεράσματα

6.1. Αποτίμηση του Έργου

Η παρούσα διατριβή πραγματεύεται την αυτοματοποιημένη και πολυδιάστατη κατηγοριοποίηση των δυναμικών δεδομένων, με στόχο την αποτελεσματικότερη διαχείριση και ανάκτησή τους. Αποτέλεσμα της ερευνητικής μας μελέτης, είναι ο σχεδιασμός της προτεινόμενης μεθοδολογίας, την οποία και εφαρμόσαμε, ώστε να είναι δυνατή η αξιολόγησή της. Το έναυσμα για την ερευνητική μελέτη αυτού του ζητήματος, αλλά και για την επιλογή των βασικών χαρακτηριστικών της πρότυπης μεθοδολογίας που προτείνουμε, είναι δύο αδιαμφισβήτητες διαπιστώσεις. Ο όγκος της πληροφορίας που διατίθεται διαδικτυακώς, σε συνδυασμό με την ποικιλομορφία της θέτουν ολοένα και μεγαλύτερες προκλήσεις στο ζήτημα της αυτοματοποιημένης διαχείρισής της, και η δυναμικότητα, που χαρακτηρίζει την πληροφορία αυτή, επιβάλλει τον συνυπολογισμό επιπλέον παραμέτρων. Δεδομένου αυτού του πλαισίου, η προτεινόμενη μεθοδολογία επιδιώκει να κατηγοριοποιήσει με τρόπο πολυδιάστατο, δηλαδή τόσο βάσει περιεχομένου όσο και βάσει δομής, δεδομένα δυναμικά, λαμβάνοντας υπόψη τον βαθμό και τον ρυθμό αλλαγής τους.

Στο πρώτο Κεφάλαιο της παρούσας εργασίας παρουσιάζεται αναλυτικά το αντικείμενο μελέτης, όπως αυτό συνοψίζεται πιο πάνω. Συμπληρωματικά, γίνεται λόγος για τη συνεισφορά της ερευνητικής μας μελέτης, καθώς και αναλυτική παρουσίαση της δομής της εν λόγω διατριβής. Στα επόμενα δύο Κεφάλαια (Κεφάλαιο 2 & 3) γίνεται εκτενής αναφορά στο θεωρητικό υπόβαθρο που πλαισιώνει τη μεθοδολογία μας. Συγκεκριμένα, στο Κεφάλαιο 2 γίνεται επισκόπηση των τεχνικών κατηγοριοποίησης, με βάση τη σύγχρονη σχετική αρθρογραφία, ενώ στο Κεφάλαιο 3 παρουσιάζονται τα βασικά χαρακτηριστικά και οι ιδιαιτερότητες του Διαδικτύου.

Στο Κεφάλαιο 4 παρουσιάζεται αναλυτικά η προτεινόμενη Μεθοδολογία, η οποία ολοκληρώνεται μέσα από τρεις ξεχωριστούς και αυτοτελείς αλγορίθμους, οι οποίοι μπορούν να λειτουργήσουν με τρόπο ανεξάρτητο, αλλά και συμπληρωματικό. Αφού γίνεται εισαγωγή στην προτεινόμενη μεθοδολογία και παρουσιάζεται η γενική

αρχιτεκτονική της, το Κεφάλαιο συνεχίζει με την αναλυτική περιγραφή των τριών αλγορίθμων που την απαρτίζουν. Ο πρώτος αλγόριθμος, αυτός της Πολυδιάστατης Κατηγοριοποίησης Σελίδων Διαδικτύου, αξιοποιώντας συγκεκριμένα δομικά και συν-κειμενικά στοιχεία των σελίδων, κατηγοριοποιεί τις σελίδες με βάση τη δομή ΚΑΙ το θέμα της καθεμιάς. Η επιλογή των στοιχείων αυτών έγινε με βάση τη βιβλιογραφία και τις δικές μας πειραματικές δοκιμές. Ο δεύτερος αλγόριθμος, Αλγόριθμος Επανακατηγοριοποίησης Σελίδων Διαδικτύου, είναι ο αλγόριθμος που, εξετάζοντας τις ίδιες σελίδες σε δύο διαφορετικές χρονικές στιγμές ως προς τα ίδια δομικά και συν-κειμενικά στοιχεία με τον πρώτο αλγόριθμο, εντοπίζει τις σελίδες που άλλαξαν, τον τρόπο που αυτές άλλαξαν (δομικά ή/και θεματικά). Έτσι, ξεχωρίζει εκείνες που χρειάζονται επανακατηγοριοποίηση, και τις στέλνει να επανακατηγοριοποιηθούν στο αντίστοιχο βήμα του Αλγορίθμου 1. Ο τρίτος αλγόριθμος, Αλγόριθμος Βελτιστοποίησης Επανακατηγοριοποίησης Σελίδων Διαδικτύου με βάση τη Συχνότητα Αλλαγής, έρχεται να βελτιώσει την απόδοση του αλγορίθμου 2, υπολογίζοντας τον ρυθμό αλλαγής των σελίδων υπό επεξεργασία, ώστε να επιλέξει και να στείλει στον αλγόριθμο 2 εκείνες μόνο τις σελίδες που εμφανίζουν ρυθμό αλλαγής τέτοιο που επιτρέπει την επανακατηγοριοποίησή τους. Κλείνοντας το Κεφάλαιο 4, γίνεται σύνοψη της Μεθοδολογίας.

Στο επόμενο κεφάλαιο, Κεφάλαιο 5, στόχος είναι η πειραματική αξιολόγηση της προτεινόμενης μεθοδολογίας. Για το σκοπό αυτό, παρουσιάζεται αναλυτικά η διαδικασία της πειραματικής εφαρμογής της Μεθοδολογίας μας, κατά την οποία, έχοντας στη διάθεσή μας ένα τυχαίο σύνολο σελίδων διαδικτύου, εφαρμόσαμε την προτεινόμενη μεθοδολογία ακολουθώντας ένα προς ένα τα βήματα που ορίζονται μέσα από τους αλγορίθμους. Στη συνέχεια, παρουσιάζονται οι μετρικές αξιολόγησης αλγορίθμων που αξιοποιήσαμε στο πλαίσιο της εργασίας μας, καθώς και αναλυτικά τα αποτελέσματα που είχαμε από αυτές, αλλά και από την πειραματική μας εφαρμογή.

6.2. Συμπεράσματα

Στο πλαίσιο της παρούσας διατριβής, παρουσιάζουμε μια πρότυπη μεθοδολογία για την πολυδιάστατη κατηγοριοποίηση των σελίδων διαδικτύου, λαμβάνοντας υπόψη τον βαθμό και τον ρυθμό αλλαγής τους. Μέσα από τους τρεις ξεχωριστούς περιγραφικούς αλγορίθμους που παρουσιάζουμε, γραμμένους σε ψευδογλώσσα, ορίζονται λεπτομερώς τα βήματα και οι διαδικασίες για την κατηγοριοποίηση των διαδικτυακών σελίδων βάσει δομής και θέματος.

Πρωταρχικό μας μέλημα είναι η ενδελεχής μελέτη προηγούμενων ερευνητικών εργασιών, που πραγματεύονται τόσο το ζήτημα της κατηγοριοποίησης κειμένων όσο και της ιδιαίτερης φύσης των σελίδων διαδικτύου, καθώς και το ολοένα και πιο απαιτητικό ζήτημα της διαχείρισης των τελευταίων. Η ερευνητική μας δουλειά μάς οδηγεί στο να κάνουμε δικές μας σκέψεις γύρω από το ζήτημα της κατηγοριοποίησης δυναμικών δεδομένων, με αποτέλεσμα το σχεδιασμό της πρότυπης μεθοδολογίας που προτείνουμε. Παράλληλα, αιτιολογώντας τις σχεδιαστικές μας αποφάσεις για κάθε βήμα ξεχωριστά, και στοιχειοθετώντας τη συλλογιστική μας πορεία για τον ορισμό του καθενός από αυτά, κάνουμε τη μεθοδολογία μας τεκμηριωμένη. Τέλος, μέσα από την πειραματική εφαρμογή σε πραγματικό σύνολο δεδομένων, η οποία μας δίνει αποτελέσματα που αξιολογούμε βάσει καθιερωμένων μετρικών, αποδεικνύουμε πως η μεθοδολογία μας είναι ορθή. Συμπερασματικά, η μεθοδολογία που προτείνουμε καταφέρνει να συνδυάσει διαφορετικές τεχνικές και εργαλεία από διαφορετικούς τομείς, τα οποία δουλεύουν συνδυαστικά και υπό προϋποθέσεις.

Ως προς την συνεισφορά της, η προτεινόμενη μεθοδολογία αποτελεί πρόταση για το συνυπολογισμό εξίσου δομικών και συν-κειμενικών στοιχείων των σελίδων διαδικτύου, με στόχο την πληρέστερη κατηγοριοποίησή τους, γεγονός που μπορεί να κάνει αποτελεσματικότερη και τη διαδικασία ανάκτησής τους. Συμπληρωματικά, η μεθοδολογία μας λαμβάνει υπόψη τη δυναμικότητα των δεδομένων αυτών, προβλέποντας και ικανοποιώντας την ανάγκη για επανακτηγοριοποίησή τους σε περιπτώσεις που κρίνεται απαραίτητο, δεδομένου του βαθμού αλλαγής τους. Τέλος, παρατηρώντας τις ίδιες σελίδες σε περιοδική βάση, η οποία προκύπτει μέσα από την

παρατήρηση της συμπεριφοράς των ίδιων των σελίδων, η μεθοδολογία μας καταφέρνει να αντιμετωπίσει τον απρόβλεπτο ρυθμό αλλαγής των διαδικτυακών δεδομένων. Ακόμα ένα χαρακτηριστικό της μεθοδολογίας μας, που την κάνει ξεχωριστή και ενισχύει την δυναμική της, είναι το γεγονός ότι μπορεί να αλλάζει, και άρα να μπορεί να βρει εφαρμογή και σε άλλα δυναμικής φύσης δεδομένα πλην των διαδικτυακών. Για παράδειγμα, κάνοντας τις απαραίτητες παραμετροποιήσεις, η μεθοδολογία μας θα μπορούσε να εφαρμοστεί σε ένα σύνολο ιατρικών γνωματεύσεων ενός νοσοκομείου, που εμπλουτίζεται συνεχώς, με σκοπό την κατηγοριοποίηση των συμπτωμάτων που παρουσιάζει κάθε ασθενής, ώστε να είναι εύκολα ανακτήσιμη η πληροφορία σχετικά με τις ασθένειες που αυτά συνδέονται και το αντίστροφο. Ο λόγος που εμείς επιλέγουμε να την διαμορφώσουμε έτσι, ώστε να μπορεί να εφαρμοστεί σε διαδικτυακά δεδομένα πάνω στα οποία κάνουμε και την πειραματική μας εφαρμογή, είναι το γεγονός ότι είναι άμεσα προσβάσιμα, ενώ διατίθενται «ελεύθερα» και δεν υπόκεινται στον κανονισμό προστασίας προσωπικών δεδομένων.

Η απόδοση της προτεινόμενης μεθοδολογίας κρίνεται αρκετά καλή, με βάση τα αποτελέσματα που παίρνουμε από την πειραματική της εφαρμογή σε ένα σύνολο 2.330 τυχαίων σελίδων διαδικτύου. Συγκεκριμένα, κατά την εφαρμογή του πρώτου αλγορίθμου (αλγόριθμος πολυδιάστατης κατηγοριοποίησης), ο αλγόριθμός μας κατάφερε να κατηγοριοποιήσει το 88% των σελίδων ως προς τη δομή τους, και το 96.80% ως προς το θέμα τους. Μάλιστα, η «βαθύτερη» δομική κατηγοριοποίηση, με δεδομένο τον τύπο των σελίδων, ήταν 100% αποτελεσματική και είχε 84% επιτυχία κατά μέσο όρο. Παράλληλα, στο πλαίσιο της θεματικής κατηγοριοποίησης, ο προτεινόμενος αλγόριθμος κατηγοριοποιεί θεματικά το 96.80% των σελίδων, και από αυτές το 83.88% ήδη από το πρώτο βήμα. Επίσης, με τη βοήθεια του αλγορίθμου επανακατηγοριοποίησης, ο οποίος, στο πλαίσιο της πειραματικής μας δοκιμής εντοπίζει αλλαγή στο 56% των σελίδων ως προς τη δομή και στο 10% των σελίδων ως προς το θέμα, αξιολόγησε πως τελικώς μόνο το 5% και 4% αντίστοιχα χρειάζεται επανακατηγοριοποίηση. Η απόδοση αυτού του αλγορίθμου, επίσης κρίνεται αρκετά καλή, με βάση της μετρικές αξιολόγησης, γεγονός που αποδεικνύει ότι η προτεινόμενη μεθοδολογία καταφέρνει να αντιμετωπίσει τη δυναμικότητα των

δεδομένων υπό επεξεργασία. Τέλος, μελετώντας τα αποτελέσματα από την εφαρμογή του Αλγορίθμου 3, σύμφωνα με τα οποία το 42% των σελίδων παρουσιάζουν αλλαγή (δομικά ή/και θεματικά), σε χρονικά διαστήματα τέτοια που υπάρχει νόημα να γίνει έλεγχος του βαθμού αλλαγής τους, φαίνεται πως η μεθοδολογία που προτείνεται καταφέρνει να αντιμετωπίσει και τον απρόβλεπτο ρυθμό αλλαγής των δεδομένων που μελετώνται στο πλαίσιο της παρούσας διατριβής.

6.3 Μελλοντικές Κατευθύνσεις

Στην παρούσα διατριβή, παρουσιάζεται μια πρότυπη μεθοδολογία πολυδιάστατης κατηγοριοποίησης σελίδων διαδικτύου, λαμβάνοντας υπόψη τόσο τη δομή όσο και το περιεχόμενο των τελευταίων. Παράλληλα, η προτεινόμενη μεθοδολογία, δεδομένης της δυναμικής φύσης των τελευταίων, καταφέρνει να αντιμετωπίσει το γεγονός ότι τα δεδομένα αλλάζουν με απρόβλεπτο τρόπο και ρυθμό.

Τα βασικά χαρακτηριστικά της προτεινόμενης μεθοδολογίας είναι τρία:

- Στηρίζεται τόσο σε δομικά όσο και σε συν-κειμενικά στοιχεία, ώστε να επιτευχθεί μια διαφορετική και πληρέστερη κατηγοριοποίηση.
- Ανιχνεύει και αξιολογεί το δυναμικό χαρακτήρα των δεδομένων, με σκοπό την επανακατηγοριοποίησή τους σε περιπτώσεις που κρίνεται απαραίτητο, και από ποιο σημείο.
- Αντιμετωπίζει τον απρόβλεπτο ρυθμό αλλαγής των δεδομένων.

Ωστόσο, οπωσδήποτε υπάρχουν περιθώρια βελτίωσης, επιπλέον παράμετροι να ελεγχθούν για την αξιολόγησή της, ενώ θα μπορούσε να αποτελέσει και αφετηρία για περαιτέρω μελέτη και έρευνα.

Αναφορικά με τις βελτιώσεις που θα μπορούσαν να γίνουν, και πιο συγκεκριμένα σχετικά με τον Αλγόριθμο 1 (Αλγόριθμος Πολυδιάστατης Κατηγοριοποίησης), το κομμάτι της δομικής κατηγοριοποίησης θα μπορούσε να βελτιωθεί με την αξιοποίηση των html elements από το sourceCode ή το viewSource των σελίδων, προσπερνώντας τις διαφορές σύνταξης που υπάρχουν μεταξύ των δεδομένων

διαδικτύου. Κάτι τέτοιο θα μπορούσε ίσως να συντομεύσει τα βήματα τόσο της δομικής όσο και της θεματικής κατηγοριοποίησης ή και να τα ενισχύει/συμπληρώνει. Για παράδειγμα, μέσα από τον html κώδικα της σελίδας, θα μπορούσε κάποιος να εντοπίζει και όρους διάδρασης που στο σώμα της σελίδας δεν έχουν τη μορφή λέξης, αλλά εικόνας ή συμβόλου.

Επίσης, στις περιπτώσεις όπου, κατά τον έλεγχο δυναμικότητας σελίδων διάδρασης εντοπίζονται νέοι όροι διάδρασης, θα μπορούσε να γίνεται αυτόματη εισαγωγή τους στο σχετικό πίνακα, ώστε και ο ίδιος να εμπλουτίζεται συνεχώς, και το «ποσοστό» αλλαγής να μην επηρεάζεται από αυτό. Έτσι, ως στοιχείο αλλαγής θα καταγράφεται μόνο η εμφάνιση/εξαφάνιση όρων διάδρασης από τη σκοπιά του «υπάρχουν/δεν υπάρχουν», και όχι από τη σκοπιά του ότι αυτοί μπορεί να άλλαξαν ή να αυξο-μειώθηκαν. Με τον τρόπο αυτό, γίνεται ακόμα πιο στοχευμένος ο υπολογισμός του βαθμού αλλαγής μιας τέτοιας σελίδας.

Πολύ βοηθητικό θα ήταν να δημιουργηθεί ένα εργαλείο υπολογισμού της αναλογίας κειμένου/(υπερ)συνδέσμων των σελίδων που να αφορά ξεκάθαρα και αποκλειστικά μόνο το κυρίως σώμα των τελευταίων, χωρίς να υπολογίζει τους (υπερ)συνδέσμους στις πλευρικές στήλες. Αυτό θα μπορούσε να βοηθήσει στον καλύτερο διαχωρισμό των σελίδων πλοήγησης από τις πληροφοριακές.

Σχετικά με τη θεματική κατηγοριοποίηση, αυτή θα μπορούσε να βελτιωθεί μέσα από την αξιοποίηση μιας λεξικογραφικής οντολογίας ή ενός θησαυρού, ώστε να είναι δυνατή η ανάθεση ενός καθιερωμένου θέματος σε κάθε σελίδα. Με αυτόν τον τρόπο, θα εντοπίζονται εύκολα και γρήγορα σελίδες ίδιας θεματολογίας. Με άλλα λόγια, χωρίς να αποτελεί στόχο η ιεραρχική κατηγοριοποίηση των σελίδων, θα μπορούσαν να αξιοποιηθούν οι ιεραρχίες για τον εντοπισμό ενός καθιερωμένου θέματος.

Επιπλέον, ενώ στο πλαίσιο της προτεινόμενης μεθοδολογίας, και για τη θεματική κατηγοριοποίηση αξιοποιούνται και οι συχνότερα εμφανιζόμενες λέξεις-κλειδιά στο περιεχόμενο μιας σελίδας (εξαιρώντας άρθρα, συνδέσμους και τερματικές λέξεις), θα μπορούσε εναλλακτικά να αξιοποιήσει κάποιος τις πρώτες λέξεις του κειμένου. Αυτό,

θα μπορούσε να έχει βάση, και να είναι αντιπροσωπευτικά του θέματος, δεδομένου ότι οι πρώτες προτάσεις ενός κειμένου μας εισάγουν στο θέμα.

Μια άλλη κατεύθυνση μελλοντικής έρευνας, με αφετηρία τον Αλγόριθμο 2 (Αλγόριθμος Επανα-κατηγοριοποίησης), θα μπορούσε να είναι η αξιοποίηση των στοιχείων της αλλαγής του μεγέθους μιας σελίδας και αυτού της τελευταίας ενημέρωσής της, προς ενίσχυσης του ελέγχου δυναμικότητας της σελίδας. Στο πλαίσιο της παρούσας διατριβής, τα στοιχεία αυτά δεν υπάρχουν για όλες τις σελίδες ή/και δεν είναι πάντα ενδεικτικά για το τελικό μας ζητούμενο.

Στο πλαίσιο βελτίωσης του ίδιου αλγορίθμου, σημασία θα είχε οι σελίδες που εμφανίζουν μεγάλο ποσοστό αλλαγής επειδή δεν είναι πια διαθέσιμες, να μπορούν να απομονωθούν, και να παύεται κάθε διαδικασία. Έτσι, δεν θα στέλνονται για επανακατηγοριοποίηση επειδή εμφανίζουν μεγάλο ποσοστό αλλαγής, και άρα θα αποφευχθεί άσκοπο «τρέξιμο» του Αλγορίθμου 2.

Τέλος, έχοντας σαν αφετηρία τον Αλγόριθμο 3 (Αλγόριθμος Βελτιστοποίησης της Επανακατηγοριοποίησης βάσει του Ρυθμού Αλλαγής των Σελίδων), ενδιαφέρον παρουσιάζει το ενδεχόμενο να ορίζονται τα Time intervals επανεξέτασης για κάθε μία σελίδα ξεχωριστά (δεδομένης της συμπεριφοράς της και κατόπιν παρατήρησης) και όχι να είναι ενιαία για όλες τις σελίδες. Έτσι, θα δοθεί στη μεθοδολογία επιπλέον ευελιξία.

Μία ακόμη πρόκληση, που αφορά την προτεινόμενη μεθοδολογία συνολικά, είναι η δοκιμή της και σε άλλα σύνολα δεδομένων, μεγαλύτερα ή/και πιο ετερογενή. Στο πλαίσιο της παρούσας εργασίας δόθηκε έμφαση στην καινοτομία και στην αποτελεσματικότητα της προτεινόμενης μεθοδολογίας, και όχι στο μέγεθος και την ποικιλομορφία του συνόλου των δεδομένων πάνω στα οποία έγινε η εφαρμογή της. Για αυτό το λόγο αποκλείσαμε εξαρχής δεδομένα με επιπλέον ιδιαίτερα και ποικίλα χαρακτηριστικά, όπως είναι τα πολυμεσικά δεδομένα (multimedia) και τα κοινωνικής δικτύωσης (social).

Συνδυαστικά, στο πλαίσιο μελλοντικής έρευνας, θα μπορούσε να δημιουργηθεί ένα σύνολο δεδομένων εκπαίδευσης για τους αλγορίθμους, το οποίο θα περιλαμβάνει αντιπροσωπευτικές σελίδες από κάθε κατηγορία, ώστε ο αλγόριθμος να εκπαιδευτεί στα βασικά χαρακτηριστικά κάθε κατηγορίας.

Στην ίδια κατεύθυνση, άξια περεταίρω έρευνας και μελέτης είναι ο έλεγχος και η αξιολόγηση της απόκρισης των αλγορίθμων μας. Στο πλαίσιο της παρούσας εργασίας, δόθηκε έμφαση και βάση στην καλή απόδοση τού κατηγοριοποιητή και η συχνότητα που αυτός θα πρέπει να «τρέχει». Ελέγχοντας και την ταχύτητα των αλγορίθμων, θα μπορούσαν να παρθούν αποφάσεις για τη βελτίωσή του. Για παράδειγμα, θα βοηθούσε στο να επιλέξει κάποιος τον τρόπο με τον οποίο θα μπορούσαν να δίνονται ορισμένα δεδομένα εισόδου, θα βλέπαμε αν η λίστα με τους όρους διάδρασης είναι καλύτερο να δίνονται υπό τη μορφή πίνακα ή βάσης δεδομένων. Παράλληλα, έχει σημασία να ελεγχθούν το εύρος δικτύου και πόρων που χρειάζονται οι αλγόριθμοι για να λειτουργήσουν. Με άλλα λόγια, έχει σημασία ο χρόνος που απαιτείται για την προ-επεξεργασία των δεδομένων, και αν είναι δυνατό να εκτελεστούν παράλληλα οι διεργασίες που ορίζονται έτσι από τον αλγόριθμο.

Τέλος, μια ακόμα παράμετρος των αλγορίθμων που κρίνεται σκόπιμο να ελεγχθεί και στη δική μας περίπτωση, είναι η πολυπλοκότητά τους. Η αξιολόγηση του στοιχείου της πολυπλοκότητας, θα μας δείξει πόση μνήμη απαιτείται για την εκτέλεση παράλληλων αιτημάτων προς διαφορετικές πηγές ή/και για την εκτέλεση ενός κοινού αιτήματος προς διαφορετικές πηγές.

Ολοκληρώνοντας την παρούσα Ενότητα, θα ήταν παράλειψη να μην γίνει αναφορά στο ενδιαφέρον που παρουσιάζει η αξιοποίηση της προτεινόμενης μεθοδολογίας και άλλου είδους δυναμικά δεδομένα πλην των διαδικτυακών. Για παράδειγμα, ενδιαφέρον παρουσιάζει η προσαρμογή της μεθοδολογίας μας και σε δεδομένα μιας μεγάλης νοσοκομειακής μονάδας ή ενός μεγάλου εκπαιδευτικού φορέα.

Αντίστοιχο ενδιαφέρον παρουσιάζει η αξιοποίηση του προτεινόμενου αλγορίθμου στο πλαίσιο του ενεργού αρχείου ενός δημόσιου οργανισμού/φορέα, ιδίως την τρέχουσα περίοδο που αναπτύσσεται και διευρύνεται συνεχώς το e-governing. Μια

άλλη διάσταση της προτεινόμενης μεθοδολογίας, θα μπορούσε να δοθεί μέσα από την αξιοποίησή της στην υπηρεσία του marketing και της διαφήμισης, αν βασιστεί σε αυτή η επεξεργασία των αγοραστικών συνηθειών των καταναλωτών που συλλέγονται μέσω των καρτών-μέλους σε μεγάλους εμπορικούς ομίλους. Από αυτή τη σκοπιά, θα έλεγε κάποιος ότι η προτεινόμενη μεθοδολογία θα μπορούσε να βρει εφαρμογή και στο πλαίσιο του intranet.

Από άλλη σκοπιά, αλλά εξίσου ενδιαφέρουσα, θα ήταν η αξιολόγηση της απόδοσης των αλγορίθμων και σε δεδομένα γραμμένα σε άλλες γλώσσες. Η παρούσα μελέτη βασίστηκε σε δεδομένα γραμμένα στην αγγλική γλώσσα, καθώς είναι η γλώσσα αναφοράς σε παγκόσμιο επίπεδο. Ωστόσο, δεδομένου ότι στο πλαίσιο της προτεινόμενης μεθοδολογίας αξιοποιούνται λεξικογραφικοί πόροι και εργαλεία επεξεργασίας φυσικής γλώσσας, έχει νόημα να δοκιμαστεί η προτεινόμενη μεθοδολογία και σε δεδομένα άλλων γλωσσών.

Παράρτημα 1: Ψευδοκώδικες αλγορίθμων προτεινόμενης μεθοδολογίας

ALGORITHM 1: Multi-Dimensional Page Classification

```
1   PROCEDURE 1: Structure-Based Classification
2     Phase 1: Page Type Recognition
3     Input: P, tokenizer, T(trans), Text-to-Link-Analyzer, (t)
4           for every P
5              look for t(trans) appearing as link
6                 if any
7                    tag P as P(transactional)
8                 else
9                    compute word tokens to links ratio (R)
10                      if R ≥ t
11                         tag P as P(informational)
12                      else
13                         tag P as P(navigational)
14                         end
15          end
16    Output: P(transactional), P(navigational), P(informational)
17    Phase 2: Layered Page Classification Given the Type
18    Input: P(transactional), P(navigational), T(corr), T(payment) D(top), LinkC, (h)
19          for every P(transactional)
20             map P(transactional) to the Table(corr)
21                for every mapping found
22                   count occurrences and tag P(transactional) with the category of max occurrence
23                else
24                   look for t(payment) appearing as link
25                      if t(payment) ≥ 1
26                         tag P(transactional) as "not-free"
27                      else
28                         tag P(transactional) as "free"
29                         end
30          for every P(navigational) starting after "http(s)://"
31             count the number of "/" in url
32                   if "/" ≥ h
33                      tag P(navigational) as "WebPage" and
34                      set the number of "/" as depth value
35                      end
36                   else
37                      tag P(navigational) as "HomePage" and
38                      map the HomePage suffix to the D(top)
39                         if there is a mapping
40                            tag HomePage with the suffix meaning
41                            end
42                         else
43                            validate url against LinkC
44                               for every valid link
45                                  if internal
46                                     set the number of (/) as depth value
47                                     end
48                                  else
49                                     send P(navigational) to Procedure1
50                                     end
51          end
52    Output: P(transactional), P(navigational), P(informational)
```

ALGORITHM 1: Multi-Dimensional Page Classification

```
PROCEDURE 2: Content-Based Classification
    Phase 1: Textual Elements Extraction
    Input: P
        for each P
            Search for anchor title in url
                if any
                    tag as "P's anchorTitle"
                    end
                Else
                    search for title in text body
                        if any
                            tag as "P's textTitle"
                            end
        end
    Output: P tagged with Textual elements
    Phase 2: Theme Detection
    Input: (P's anchorTitle), (P's textTitle), WebPage Word Counter, PoS-Tagger, Parser, WordNet, lemmatizer, (TF*IDF), (n), WP contents, WP articles
        for each P look for common terms between P's anchorTitle and P's textTitle
          if found
            use common terms as the thematic term(-s) to tag P and map thematic(s) term(s) to WP contents
              for every mapping found
                count occurrences and tag P with the category of max occurrence
                end
              else
                search for WP article titled with the wider thematic term
                map article's categories to WP contents begging from the first one
                stop when a mapping is found and tag P with the category
                end
          else
            PoS-tag and lemmatize P's text and extract the first n-appearing keywords
            check for overlapping terms between P's keywords and (P's anchor title and P's text title)
              if found
                use overlapping terms as the thematic term(-s) to tag P and map thematic(s) term(s) to WP contents
                  for every mapping found
                    count occurrences and tag P with the category of max occurrence
                    end
                  else
                    search for WP article titled with the wider thematic term
                    map article's categories to WP contents begging from the first one
                    stop when a mapping is found and tag P with the category
                    end
          else
            map P's first n-appearing keywords to WordNet and look for common senses between P's keywords and (P's anchor title and P's text title)
              if found
                use terms of common senses as the thematic term(-s) to tag P and map thematic(s) term(s) to WP contents
                  for every mapping found
                    count occurrences and tag P with the category of max occurrence
                    end
                  else
                    search for WP article titled with the wider thematic term
                    map article's categories to WP contents begging from the first one
                    stop when a mapping is found and tag P with the category
                    end
            else
              tag P as unknown category (Punknown)
              end
        end
```

58 Output: Thematically classified *P*
59 Output: Multi-Dimensionally classified *P*

ALGORITHM 2: Re-Classification based on Change Detection

```
Input: P(class, T), P'(unclass, T')
   PROCEDURE 1: Re-Classification Decision based on Textual Changes
   Input: (E(t) ∈ P), (E(t) ∈ P'), smlrtMetric, (m), (z)
      for each pair of (Pi ∈ P(class,T), (P'i ∈ P'(unclass, T'))
         compute sim(Pi, P'i)
            if sim(Pi, P'i) ≥ m
               tag P'i as thematically unchanged and classify P'i to the category of Pi
               end
            else
               tag P'i as thematically changed and
                  compare (E(t) ∈ P'i) with (E(t) ∈ Pi)
                  count ((E(t) ∈ P'i) ≠ (E(t) ∈ Pi))
                     if ((E(t) ∈ P'i) ≠ (E(t) ∈ Pi)) ≤ z
                        go to Algorithm2Procedure2
                        end
                     else
                        send P'i to Algorithm1Procedure2
                        end
      end
   Output: thematically unchanged pages P' over time T'
   Procedure2: Re-Classification Decision based on Structural Changes
   Input: (E(s) ∈ P), (E(s) ∈ P'), smlrtMetric, (z)
      for each pair of (Pi ∈ P(class,T), (P'i ∈ P'(unclass, T'))
         compare (E(s) ∈ Pi) with (E(s) ∈ P'i) and
         count ((E(s) ∈ Pi) ≠ (E(s) ∈ P'i))
            if ((E(s) ∈ Pi) ≠ (E(s) ∈ P'i)) ≤ z
               tag P'i as structurally unchanged and classify P'i to the category of Pi
               end
            else
               send P'i to Algorithm1Procedure1
               end
   Output: structurally unchanged pages P' over time T'
Output: P'(ReClass, T'), thematically unchanged pages P' over time T', structurally unchanged pages P' over time T'
```

ALGORITHM 3: Optimized Re-Classification based on Change's Frequency Detection

```
1   Input: (P(class, T)), ((E(t) U E(s)) ∈ P(class, T)), P' ⊆ (P'(re-class, T'), ((E(t) U E(s)) ∈ P'(reClass,T')), MaxFreqChange,
    MinFreqChange, Timer
2       when Algorithm 2 initializes, record Ts
3       for every pair of ((Pi ∈ P(class, T), (P'i ∈ P'(re-class, T'))
4           set Timer
5           while ((E(t) U E(s)) ∈ Pi(class, T)) ≠ ((E(t) U E(s)) ∈ P'i(re-class,T')), record Ts
6               if Ts ≥ MaxFreqChange
7                   tag P'i as HighlyChanging Page and keep it in a secondary Index
8                   end
9               else
10                  if Ts ≤ MinFreqChange
11                      tag P'i as RarelyChanging Page and keep it in a secondary Index
12                      end
13                  else
14                      tag P'i as RegularlyChanging Page and send it to Algorithm2
15                      end
16      end
17  Output: Selection of Pages that need periodical Re-Classification
```

Παράρτημα 2: Ορολογία και λεκτικά αλγορίθμων (ελληνικά)

(E(s) ∈ P): λίστα τιμών δομικών στοιχείων των σελίδων υπό επεξεργασία σε χρόνο T

(E(s) ∈ P'): λίστα τιμών δομικών στοιχείων των σελίδων υπό επεξεργασία σε χρόνο T'

(E(t) ∈ P): λίστα τιμών κειμενικών στοιχείων των σελίδων υπό επεξεργασία σε χρόνο T

(E(t) ∈ P'): λίστα τιμών κειμενικών στοιχείων των σελίδων υπό επεξεργασία σε χρόνο T'

(E(t) U E(s)) ∈ P(class, T): λίστα τιμών κειμενικών και δομικών στοιχείων των σελίδων υπό επεξεργασία, κατά την πρώτη τους κατηγοριοποίηση σε χρόνο T

(E(t) U E(s)) ∈ P'(re-class,T'): λίστα τιμών κειμενικών και δομικών στοιχείων των σελίδων υπό επεξεργασία σε χρόνο T'

(h): όριο για την ανίχνευση των ιστοσελίδων

(z): όριο για την ανίχνευση των δομικά όμοιων στιγμιοτύπων της ίδιας σελίδας

(n): όριο για την εξαγωγή των λέξεων-κλειδιά

(m): όριο για την ανίχνευση των θεματικά όμοιων στιγμιοτύπων της ίδιας σελίδας

(t): όριο για την ανίχνευση των πληροφοριακών σελίδων

(TF*IDF): συντομογραφία του τύπου «term frequency–inverse document frequency», αριθμητική έκφραση στατιστικής για τον υπολογισμό της σπουδαιότητας μιας λέξης στο πλαίσιο ενός εγγράφου/κειμένου ή μιας συλλογής εγγράφων/κειμένων

(Ts): καταγεγραμμένη χρονική στιγμή

D(top): λίστα με τους τομείς ανώτατου επιπέδου, αναφορικά με τα επιθέματα των URL's

Lemmatizer: γλωσσικό εργαλείο που δέχεται ως είσοδο έναν οποιοδήποτε λεκτικό τύπο και επιστρέφει το λημματικό τύπο στον οποίο αντιστοιχεί, π.χ. για τον τύπο *κατέστη* επιστρέφει το λημματικό τύπο *καθιστώ*

LinkC: εργαλείο επεξεργασίας συνδέσμων που περιέχονται σε μια σελίδα διαδικτύου

MaxFreqChange: μέγιστη επιτρεπτή από τον αλγόριθμο συχνότητα αλλαγής των σελίδων

MinFreqChange: ελάχιστη επιτρεπτή από τον αλγόριθμο συχνότητα αλλαγής των σελίδων

P' ⊆ (P'(re-class, T'): υποσύνολο στιγμιοτύπων σελίδων υπό επεξεργασία σε χρόνο T', που χρειάζεται να επανακατηγοριοποιηθούν στον χρόνο T'

P(class, T): σελίδες κατηγοριοποιημένες από τον Αλγόριθμο 1 σε χρόνο T

P'(class, T'): σελίδες κατηγοριοποιημένες από τον Αλγόριθμο 1 σε χρόνο T'

P(navigational): σελίδες πλοήγησης

P(transactional): σελίδες διάδρασης

P(informational): πληροφοριακές σελίδες

P: σελίδες προς κατηγοριοποίηση

P'(ReClass, T'): σελίδες που μεταβάλλονται κειμενικά ή/και δομικά μέσα στο χρόνο και χρειάζεται να επανακατηγοριοποιηθούν

(P's anchorTitle): το στοιχείο *anchor title* της σελίδας

(P's textTitle): το στοιχείο *τίτλος* της σελίδας

Parser: εργαλείο που «σπάει» τα συστατικά μέρη μιας περιόδου φυσικής γλώσσας, και τα αναλύει με βάση τον συντακτικό τους ρόλο

PoS-Tagger: λογισμικό που «διαβάζει» κείμενο γραμμένο σε φυσική γλώσσα και επισημειώνει κάθε όρο με το μέρος του λόγου που είναι

smlrtMetric: μετρική ομοιότητας

T(corr): πίνακας αντιστοίχισης όρων διάδρασης με τους σκοπούς που εξυπηρετούν

T(payment): πίνακας με τους όρους που δηλώνουν συναλλαγή

t(payment): όρος που δηλώνει συναλλαγή

T(trans): πίνακας με τους όρους που υποδηλώνουν διάδραση μεταξύ χρήστη και σελίδας διαδικτύου

t(trans): όροι που υποδηλώνουν διάδραση μεταξύ χρήστη και σελίδας διαδικτύου

Text-to-Link-Analyzer: εργαλείο για τον υπολογισμό της αναλογίας μεταξύ λεκτικών όρων και (υπερ)συνδέσμων

Timer: timer that calculates the time based on a defined formula (e.g. if t is the time when a webpage is first examined, the timer will calculate every t_i time instance that we want to re-examine the same page, according to the formula $t_i=(t_i-1)*$constant. Each time of re-examination is the multiplication product of its preceding with the constant defined.).

Tokenizer: εργαλείο για τη διαίρεση ενός κειμένου σε τμήματα με νόημα, το μικρότερο τμήμα (συμβολοσειρά) με νόημα είναι μια λέξη

WebPage Word Counter: εργαλείο για την εξαγωγή λέξεων-κλειδιών

WordNet: ιεραρχικά οργανωμένο δίκτυο σημασιολογικών λημμάτων

Terminology of Algorithms (english)

(E(s) ∈ P): list of P structural elements

(E(s) ∈ P'): list of P' structural elements

(E(t) ∈ P): list of P textual elements

(E(t) ∈ P'): list of P' textual elements

(E(t) U E(s)) ∈ P(class, T): list of P(class, T) textual and structural elements

(E(t) U E(s)) ∈ P'(re-class,T'): textual and structural elements of P'(re-class,T')

(h): threshold for homepages' detection

(z): threshold for structurally unchanged pages' detection

(n): threshold for keywords

(m): threshold for the theamatically unchanged pages' detection

(t): threshold for the informational pages' detection

(TF*IDF): short for "term frequency–inverse document frequency", is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

(Ts): time stamp

D(top): list of web top-level domains

Lemmatizer: tool that groups inflected forms together as a single base form

LinkC: link counter (tool)

MaxFreqChange: maximum frequency "allowed" by the algorithm for changes to webpages

MinFreqChange: minimum frequency "allowed" by the algorithm for changes to webpages

P' ⊆ (P'(re-class, T'): subset of P' that have been ReClassified based on Algorithm2 at time T'

P(class, T): pages classified from Algorithm1 at time T

P'(class, T'): pages classified from Algorithm1 at time T'

P(navigational): navigational pages

P(transactional): transactional pages

P(informational): informational pages

P: pages for classification

P'(ReClass, T'): textually or/and structurally changed pages P' over time that need to be ReClassified

(P's anchorTitle): page's anchor title

(P's textTitle): page's text title

Parser: compiler or interpreter component that breaks data into smaller elements for easy translation into another language.

PoS-Tagger: software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc.

smlrtMetric: similarity metric

T(corr): table with transactions' correspondences

T(payment): table with Payment terms [t(payment)]

t(payment): payment terms

T(trans): table with transactional terms

t(trans): transactional terms

Text-to-Link-Analyzer: tool for the calculation of WordTokens2Links Ratio

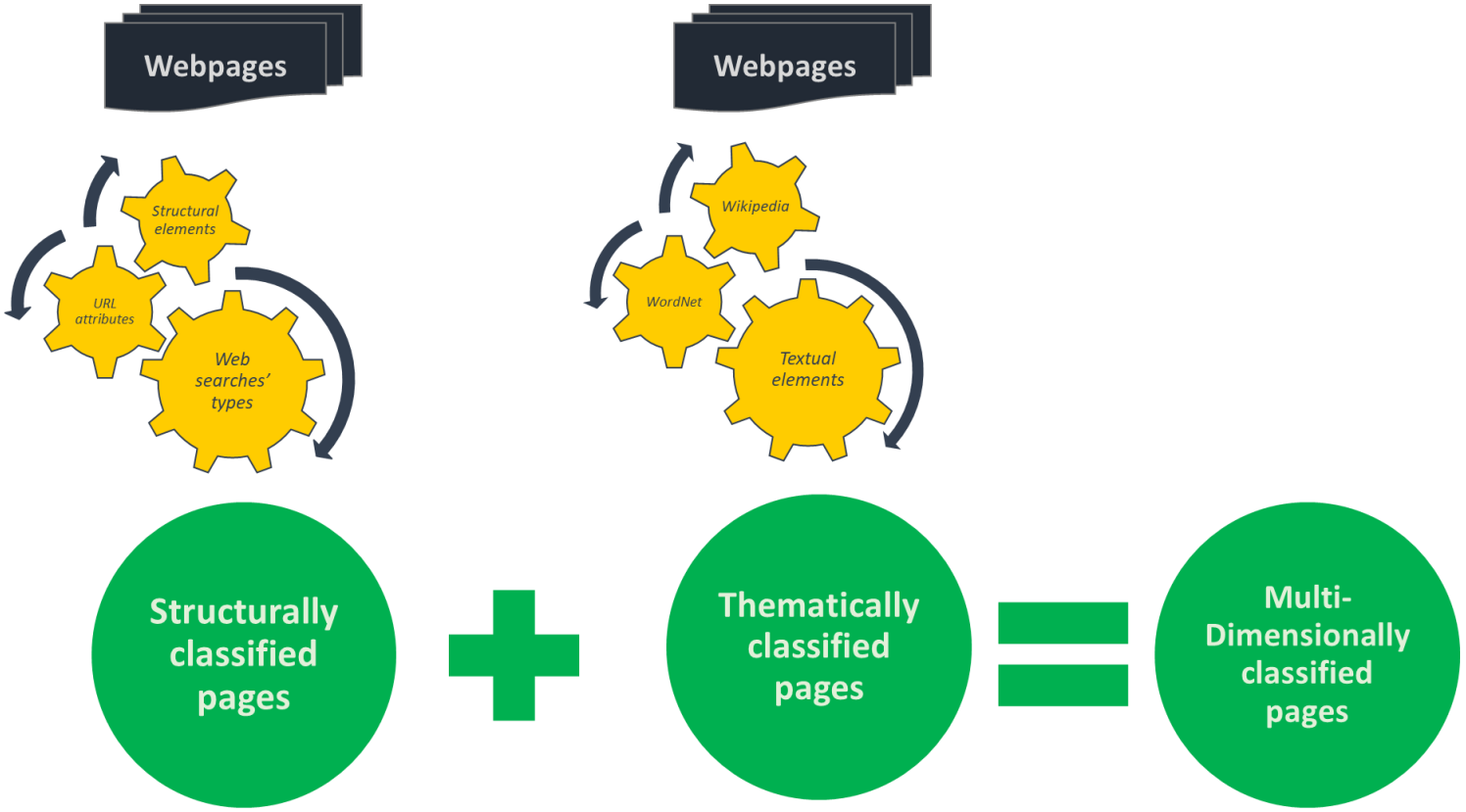
Timer: timer that calculates the time based on a defined formula (e.g. if t is the time when a webpage is first examined, the timer will calculate every t_i time instance that we want to re-examine the same page, according to the formula $t_i=(t_i-1)*$constant. Each time of re-examination is the multiplication product of its preceding with the constant defined.).

Tokenizer: tool for the tokenization of the text

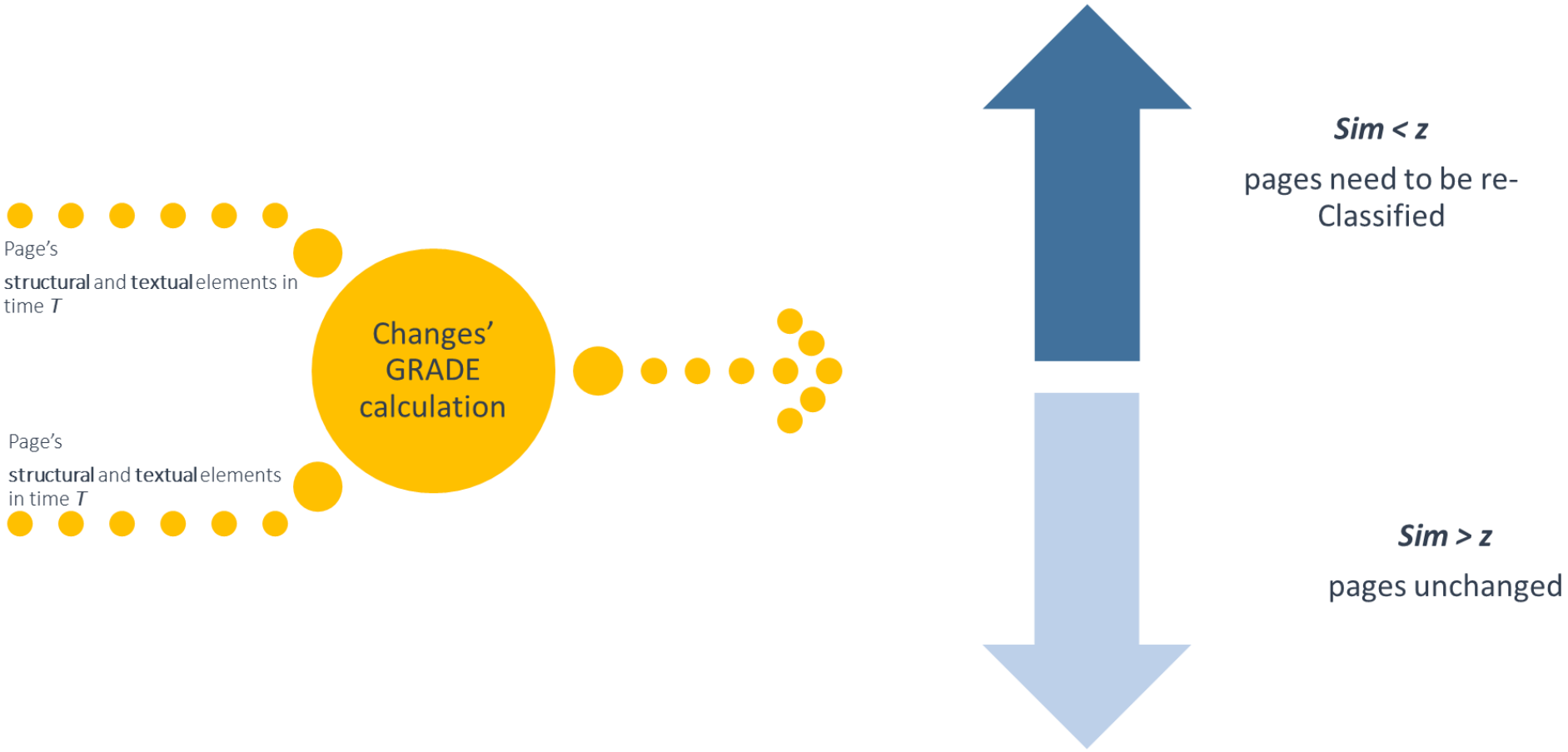
WebPage Word Counter: tool for keywords' extraction

WordNet: hierarchically organized dictionary

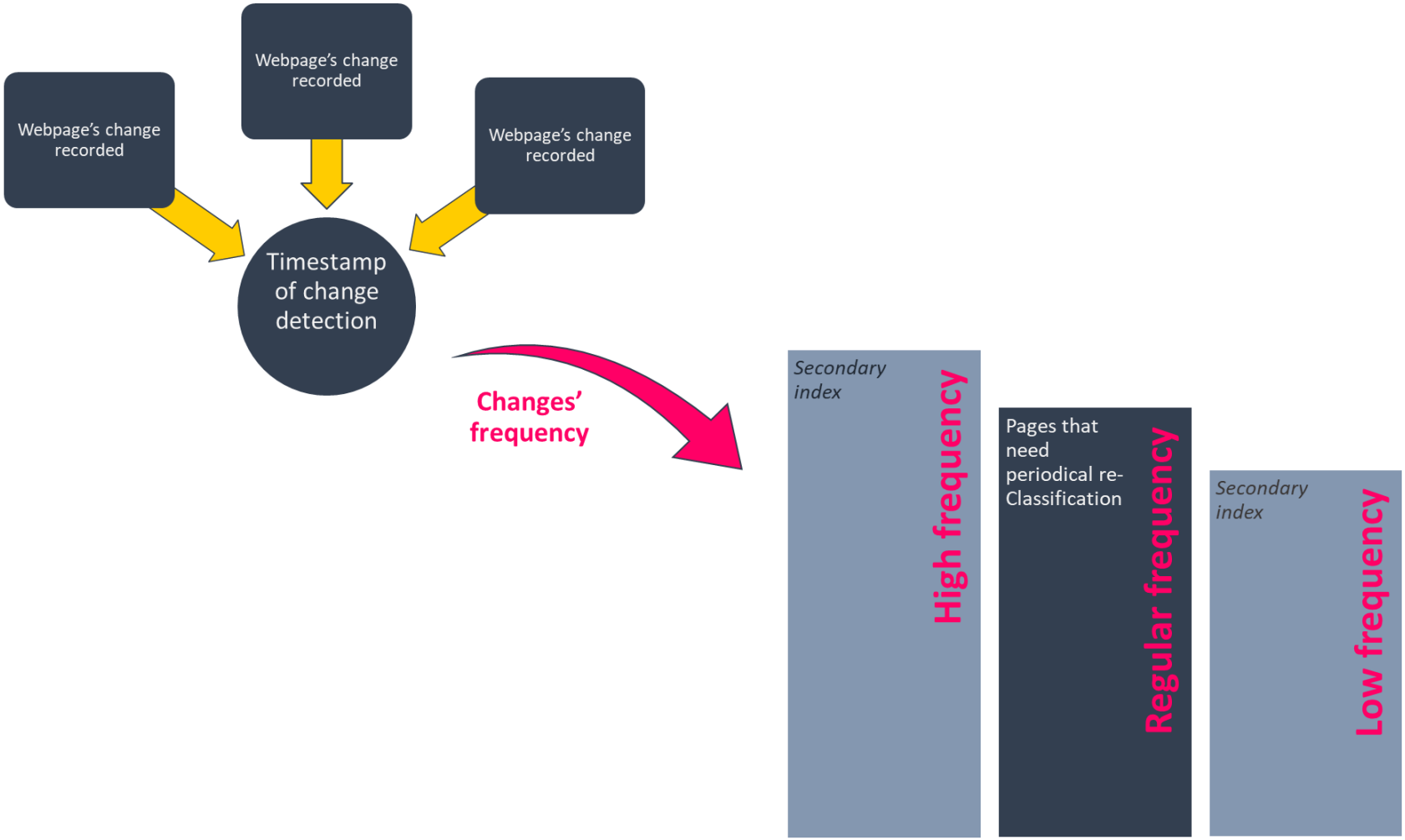
Παράρτημα 3: Σχηματική απεικόνιση προτεινόμενης μεθοδολογίας



Εικόνα 11: Σχηματική απεικόνιση διαδικασιών Αλγορίθμου Πολυδιάστατης Κατηγοριοποίησης Σελίδων Διαδικτύου



Εικόνα 12: Σχηματική απεικόνιση διαδικασιών Αλγορίθμου Επανα-κατηγοριοποίησης Σελίδων Διαδικτύου με βάση τον Βαθμό Αλλαγής



Εικόνα 13: Σχηματική απεικόνιση διαδικασιών Αλγορίθμου Βελτιστοποίησης Επανακατηγοριοποίησης Σελίδων Διαδικτύου με βάση τον Ρυθμό Αλλαγής

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Abdelali, A., Cowie, J., Farwell, D., Ogden, B., & Helmreich, S. (2003, June). Cross-language information retrieval using ontology. In *Proc. of the Conference TALN 2003* (pp. 72-86).
2. Adami, G., Avesani, P., & Sona, D. (2003, November). Clustering documents in a web directory. In *Proceedings of the 5th ACM international workshop on Web information and data management* (pp. 66-73).
3. Allahyari, M., & Kochut, K. (2016, February). Semantic tagging using topic models exploiting Wikipedia category network. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)* (pp. 63-70). IEEE.
4. Alom, B. M. (2016). WEB DATA MINING: VIEWS OF CRIMINAL ACTIVITIES . *European Journal of Computer Science and Information Technology*, *4*(4), 28-40.
5. Amitay, E., Carmel, D., Darlow, A., Lempel, R., & Soffer, A. (2003). The connectivity sonar: detecting site functionality by structural patterns. In *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia* (pp. 38-47).
6. Arya, C., & Dwivedi, S. K. (2016, October). News web page classification using url content and structure attributes. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)* (pp. 317-322). IEEE.
7. Asirvatham, A. P., & Ravi, K. K. (2001). Web page classification based on document structure. In *IEEE National Convention*.
8. Avramidis, D., Kyriakopoulou, M., Tzagarakis, M., Stamou, S., & Christodoulakis, D. (2002, August). Approaching wordnets through a structural point of view. In *International Symposium on Metainformatics* (pp. 49-57). Springer, Berlin, Heidelberg.
9. Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). New York: ACM press.
10. Beckett, D., & McBride, B. (2004). RDF/XML syntax specification (revised). *W3C recommendation*, *10*(2.3).

11. Berners-Lee, T. (2006). Linked data-design issues. *http://www. w3. org/DesignIssues/LinkedData. html*.
12. Berners-Lee, T. (2010). The original proposal of the WWW, HTMLized. 1998.
13. Berners-Lee, T., Fielding, R., & Frystyk, H. (1996). Hypertext transfer protocol--HTTP/1.0.
14. Berners-Lee, T., Fielding, R., & Masinter, L. (1998). Uniform resource identifiers (URI): Generic syntax.
15. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american*, *284*(5), 34-43.
16. Bhimavaram, S., & Govindarajulu, P. (2015). An Enhanced Approach for Ontology based Classification in Semantic Web Technology. International Journal of Advanced Research in Computer and Communication Engineering, 4, 2.
17. Bizer, C., Heath, T., & Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts* (pp. 205-227). IGI Global.
18. Brooks, T. A. (2010). World wide web consortium (W3C). In *Encyclopedia of library and information sciences* (pp. 5695-5699).
19. Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science journal*, *14*.
20. Chandarana, P., & Vijayalakshmi, M. (2014, April). Big data analytics frameworks. In *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)* (pp. 430-434). IEEE.
21. Ciglan, M., Laclavík, M., & Dorman, A. (2014). Reusing knowledge hidden in wikipedia for scalable text categorization. In *Proceedings of WSDM Workshops: WSCBD, WSDM* (Vol. 14).
22. De Luca, E. W., Nürnberger, A., & Von-Guericke, O. (2004, June). Ontology-based semantic online classification of documents: Supporting users in searching the web. In *Proc. of the European Symposium on Intelligent Technologies (EUNITE 2004), Aachen*.

23. Edmonds, A., White, R. W., Morris, D., & Drucker, S. M. (2019). Instrumenting the dynamic web. *Journal of Web Engineering*, *6*.
24. Eriksson, T. (2013). Automatic web page categorizationusing text classication methods.
25. Fürnkranz, J. (1999). Exploiting structural information for text classification on the WWW. In *International Symposium on Intelligent Data Analysis* (pp. 487-497). Springer, Berlin, Heidelberg.
26. Gali, N., & Fränti, P. (2016, April). Content-based Title Extraction from Web Page. In *WEBIST (2)* (pp. 204-210).
27. García-Pedrajas, N., Del Castillo, J. A. R., & Cerruela-García, G. (2015). A proposal for local $ k $ values for $ k $-nearest neighbor rule. *IEEE transactions on neural networks and learning systems*, *28*(2), 470-475.
28. Glover, E. J., Tsioutsiouliklis, K., Lawrence, S., Pennock, D. M., & Flake, G. W. (2002, May). Using web structure for classifying and describing web pages. In *Proceedings of the 11th international conference on World Wide Web* (pp. 562-569).
29. Golub, K., & Ardö, A. (2005). Importance of HTML structural elements and metadata in automated subject classification. In *International Conference on Theory and Practice of Digital Libraries* (pp. 368-378). Springer, Berlin, Heidelberg.
30. Golub, K., & Ardö, A. (2005, September). Importance of HTML structural elements and metadata in automated subject classification. In *International Conference on Theory and Practice of Digital Libraries* (pp. 368-378). Springer, Berlin, Heidelberg.
31. Gonzalo, J., Verdejo, F., Chugur, I., & Cigarran, J. (1998). Indexing with WordNet synsets can improve text retrieval.
32. Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199-220.
33. Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 986-996). Springer, Berlin, Heidelberg.

34. Guo, N., He, Y., Yan, C., Liu, L., & Wang, C. (2016, December). Multi-level topical text categorization with wikipedia. In *Proceedings of the 9th International Conference on Utility and Cloud Computing* (pp. 343-352).
35. Hall, W., & Tiropanis, T. (2012). Web evolution and Web science. *Computer Networks*, *56*(18), 3859-3865.
36. Hashemi, M. (2020). Web page classification: a survey of perspectives, gaps, and future directions. *Multimedia Tools and Applications*, 1-25.
37. Janik, M., & Kochut, K. J. (2008, August). Wikipedia in action: Ontological knowledge in text categorization. In *2008 IEEE International Conference on Semantic Computing* (pp. 268-275). IEEE.
38. Jansen, B. J., Booth, D. L., & Spink, A. (2007). Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web* (pp. 1149-1150).
39. Jansen, B. J., Booth, D. L., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management*, *44*(3), 1251-1266.
40. Jatowt, A., & Ishizuka, M. (2004, September). Summarization of dynamic content in web collections. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 245-254). Springer, Berlin, Heidelberg.
41. Kan, M. Y. (2004). Web page classification without the web page. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters* (pp. 262-263).
42. Kan, M. Y., & Thi, H. O. N. (2005). Fast webpage classification using URL features. In *Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 325-326).
43. Kim, W., Jeong, O. R., & Lee, S. W. (2010). On social Web sites. *Information systems*, *35*(2), 215-236.
44. Labrou, Y., & Finin, T. (1999, November). Yahoo! as an ontology: using Yahoo! categories to describe documents. In *Proceedings of the eighth international conference on Information and knowledge managemen* (pp. 180-187).

45. Li, H., Xu, Z., Li, T., Sun, G., & Choo, K. K. R. (2017). An optimized approach for massive web page classification using entity similarity based on semantic network. Future Generation Computer Systems, 76, 510-518.
46. Lindemann, C., & Littig, L. (2006). Coarse-grained classification of web sites by their structural properties. In Proceedings of the 8th annual ACM international workshop on Web information and data management (pp. 35-42).
47. Mallawaarachchi, V., Meegahapola, L., Madhushanka, R., Heshan, E., Meedeniya, D., & Jayarathna, S. (2020). Change detection and notification of web pages: A survey. *ACM Computing Surveys (CSUR)*, *53*(1), 1-35.
48. Matosevic, G. (2015). Using anchor text to improve web page title in process of search engine optimization. In *Central European Conference on Information and Intelligent Systems* (p. 173). Faculty of Organization and Informatics Varazdin.
49. Medeiros, J. F., Nunes, B. P., Siqueira, S. W. M., & Leme, L. A. P. P. (2018, June). Tagtheweb: Using wikipedia categories to automatically categorize resources on the web. In *European Semantic Web Conference* (pp. 153-157). Springer, Cham.
50. Meegahapola, L., Alwis, R., Nimalarathna, E., Mallawaarachchi, V., Meedeniya, D., & Jayarathna, S. (2017, September). Detection of change frequency in web pages to optimize server-based scheduling. In *2017 Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer)* (pp. 1-7). IEEE.
51. Mörch, C. M., Cote, L. P., Corthesy-Blondin, L., Plourde-Léveillé, L., Dargis, L., & Mishara, B. L. (2018). The Darknet and suicide. *Journal of affective disorders*, *241*, 127-132.
52. Moro, A., & Navigli, R. (2012, October). WiSeNet: Building a Wikipedia-based semantic network with ontologized relations. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1672-1676).
53. Murugesan, S. (Ed.). (2009). *Handbook of research on Web 2.0, 3.0, and X. 0: technologies, business, and social applications: Technologies, business, and social applications*. IGI Global.

54. Neches, R., Fikes, R. E., Finin, T., Gruber, T., Patil, R., Senator, T., & Swartout, W. R. (1991). Enabling technology for knowledge sharing. *AI magazine*, *12*(3), 36-36.
55. Niarou, M., & Stamou, S. (2012). Exploring lexicographic ontologies for hierarchically organizing the Greek Wikipedia articles. *J. Digit. Inf. Manag.*, *10*(3), 157-167.
56. Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental journal of computer science & technology*, *8*(1), 13-19.
57. Olla, P., & Qureshi, E. (2010). Integration of Web 2.0 Collaboration Tools into Education: Lessons Learned. In *Handbook of Research on Web 2.0, 3.0, and X. 0: Technologies, Business, and Social Applications* (pp. 522-538). IGI Global.
58. Ostermaier, B., Römer, K., Mattern, F., Fahrmair, M., & Kellerer, W. (2010). A real-time search engine for the web of things. In *2010 Internet of Things (IOT)* (pp. 1-8). IEEE.
59. Ostermaier, B., Römer, K., Mattern, F., Fahrmair, M., & Kellerer, W. (2010). A real-time search engine for the web of things. In *2010 Internet of Things (IOT)* (pp. 1-8). IEEE.
60. Peng, X., & Choi, B. (2002). Automatic web page classification in a dynamic and hierarchical way. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* (pp. 386-393). IEEE.
61. Pierre, J. M. (2001). On the automated classification of web sites.
62. Qi, X., & Davison, B. D. (2006). Knowing a web page by the company it keeps. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 228-237).
63. Qi, X., & Davison, B. D. (2009). Web page classification: Features and algorithms. *ACM computing surveys (CSUR)*, *41*(2), 1-31.
64. Qi, X., & Davison, B. D. (2009). Web page classification: Features and algorithms. *ACM computing surveys (CSUR)*, *41*(2), 1-31.
65. Rajalakshmi, R., & Aravindan, C. (2018). A Naive Bayes approach for URL classification with supervised feature selection and rejection framework. *Computational Intelligence*, *34*(1), 363-396.

66. Riboni, D. (2002). *Feature selection for web page classification*. na.
67. Rokach, L., & Maimon, O. Z. (2008). *Data mining with decision trees: theory and applications* (Vol. 69). World scientific.
68. S. Shinde, P. Joeg, and S. Vanjale, "Web Document Classification using Support Vector Machine," in 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Sep. 2017, pp. 688–691, doi: 10.1109/CTCEEC.2017.8455102.
69. Saleh, A. I., Al Rahmawy, M. F., & Abulwafa, A. E. (2017). A semantic based Web page classification strategy using multi-layered domain ontology. *World Wide Web*, *20*(5), 939-993.
70. Salton, G., & Yang, C. S. (1973). *On the specification of term values in automatic indexing*. Cornell University.
71. Sara-Meshkizadeh, D., & Masoud-Rahmani, A. (2010). Webpage classification based on compound of using HTML features & URL features and features of sibling pages. *International Journal of Advancements in Computing Technology*, *2*(4), 36-46.
72. Sara-Meshkizadeh, D., & Masoud-Rahmani, A. (2010). Webpage classification based on compound of using HTML features & URL features and features of sibling pages. *International Journal of Advancements in Computing Technology*, *2*(4), 36-46.
73. Schönhofen, P. (2009). Identifying document topics using the Wikipedia category network. *Web Intelligence and Agent Systems: An International Journal*, *7*(2), 195-207.
74. Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.
75. Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In Emerging technology in modelling and graphics (pp. 99-111). Springer, Singapore.
76. Shu, K., Bernard, H. R., & Liu, H. (2019). Studying fake news via network analysis: detection and mitigation. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining* (pp. 43-65). Springer, Cham.

77. Stamou, S., Ntoulas, A., Krikos, V., Kokosis, P., & Christodoulakis, D. (2006). Classifying web data in directory structures. In *Asia-Pacific Web Conference* (pp. 238-249). Springer, Berlin, Heidelberg.
78. Suchecki, K., Salah, A. A. A., Gao, C., & Scharnhorst, A. (2012). Evolution of wikipedia's category structure. *Advances in complex systems*, *15*(supp01), 1250068.
79. Taghva, K., Borsack, J., Coombs, J., Condit, A., Lumos, S., & Nartker, T. (2003, April). Ontology-based classification of email. In *Proceedings ITCC 2003. International Conference on Information Technology: Coding and Computing* (pp. 194-198). IEEE.
80. Tan, S. (2005). Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications*,*28*(4), 667-671.
81. Tenenboim, L., Shapira, B., & Shoval, P. (2008). Ontology-based classification of news in an electronic newspaper.
82. Tran, N. K., Sheng, Q. Z., Babar, M. A., & Yao, L. (2017). Searching the web of things: State of the art, challenges, and solutions. *ACM Computing Surveys (CSUR)*, *50*(4), 1-34.
83. Utard, H., & Fürnkranz, J. (2005). Link-local features for hypertext classification. In *Semantics, web and mining* (pp. 51-64). Springer, Berlin, Heidelberg.
84. Vogrinčič, S., & Bosnić, Z. (2011). Ontology-based multi-label classification of economic articles. *Computer Science and Information Systems*, *8*(1), 101-119.
85. Wijewickrema, C. M., & Gamage, R. (2013). An ontology based fully automatic document classification system using an existing semi-automatic system.
86. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.
87. Zanasi, A. (Ed.). (2007). *Text mining and its applications to intelligence, CRM and knowledge management* (Vol. 7). Wit Press.
88. Zeng, D., Guo, S., & Cheng, Z. (2011). The web of things: A survey. *JCM*, *6*(6), 424-438.

89. Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(3), 1-19.
90. Zhou, X., & Zafarani, R. (2018). Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, *2*.
91. Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019, January). Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the twelfth ACM international conference on web search and data mining* (pp. 836-837).
92. Asirvatham, A. P., & Ravi, K. K. (2001, December). Web page classification based on document structure. In *IEEE National Convention*.

ΔΙΑΔΙΚΤΥΑΚΕΣ ΑΝΑΦΟΡΕΣ

I. http://www.thinkmind.org/index.php?view=article&articleid=web_2020_1_20_40017 (Ημ. Τελ. Επίσκεψης 25/11/2021)

II. https://en.wikipedia.org/wiki/List_of_Internet_top-level_domains (Ημ. Τελ. Επίσκεψης 25/11/2021)

III. https://en.wikipedia.org/wiki/Wikipedia:Categorization#Articles (Ημ. Τελ. Επίσκεψης 25/11/2021)

IV. https://sonovabitc.win/analyze.php (Ημ. Τελ. Επίσκεψης 25/11/2021)

V. https://visualping.io/ (Ημ. Τελ. Επίσκεψης 25/11/2021)

VI. https://wordcounter.net/website-word-count (Ημ. Τελ. Επίσκεψης 25/11/2021)

VII. https://www.w3.org/Addressing/URL/url-spec.txt) (Ημ. Τελ. Επίσκεψης 25/11/2021)

VIII. https://www.w3schools.com/html/html_attributes.asp (Ημ. Τελ. Επίσκεψης 25/11/2021)

IX. https://www.w3schools.com/html/html_headings.asp (Ημ. Τελ. Επίσκεψης 25/11/2021)

X. https://www.wachete.com/ (Ημ. Τελ. Επίσκεψης 25/11/2021)