Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Μοντελοποίηση, Ανάλυση και Διαφοροποιημένη Ανάκτηση Νομικής Πληροφορίας

Διδακτορική Διατριβή

του

Κόνιαρη Μάριου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π. (1999)
Μεταπτυχιακό Δίπλωμα Ειδίκευσης στη Διοίκηση Επιχειρήσεων (MBA) Ο.Π.Α - Ε.Μ.Π. (2008)
Μεταπτυχιακό Δίπλωμα Ειδίκευσης στη Διαχείριση Τεχνικών Έργων Ε.Α.Π. (2012)

Αθήνα, Ιούλιος 2017

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Μοντελοποίηση, Ανάλυση και Διαφοροποιημένη Ανάκτηση Νομικής Πληροφορίας

Διδακτορική Διατριβή

του

Κόνιαρη Μάριου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π. (1999)
Μεταπτυχιακό Δίπλωμα Ειδίκευσης στη Διοίκηση Επιχειρήσεων (MBA) Ο.Π.Α - Ε.Μ.Π. (2008)
Μεταπτυχιακό Δίπλωμα Ειδίκευσης στη Διαχείριση Τεχνικών Έργων Ε.Α.Π. (2012)

Συμβουλευτική Επιτροπή: Ι. Βασιλείου
Τ. Σελλής
Κ. Κοντογιάννης

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 26η Ιουλίου 2017.

Ι. Βασιλείου	Τ. Σελλής	Κ. Κοντογιάννης
Ομ. Καθ. ΕΜΠ	Καθ. SWINBURNE	Καθ. ΕΜΠ
Σ. Κόλλιας	Ι. Αναγνωστόπουλος	Π. Τσανάκας
Καθ. ΕΜΠ	Αναπλ. Καθ. Παν. Θεσ.	Καθ. ΕΜΠ
	Γ. Στάμου	
	Αναπλ. Καθ. ΕΜΠ	

Αθήνα, Ιούλιος 2017

.....

Κονιαρης Μαριος
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Πρόλογος

Η παρούσα διατριβή εκπληρώνει τις απαιτήσεις για την απόκτηση διπλώματος του Διδάκτορα της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, του Εθνικού Μετσόβιου Πολυτεχνείου. Η διατριβή παρουσιάζει μεθόδους μοντελοποίησης, ανάλυσης και διαφοροποιημένης ανάκτησης νομικής πληροφορίας και πραγματοποιήθηκε στο Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων. Η παρούσα διατριβή αποτελεί έργο του συγγραφέα, ωστόσο, για να ολοκληρωθεί αυτό το ταξίδι απαιτήθηκε η βοήθεια πολλών ανθρώπων, άμεσα ή έμμεσα σχετιζόμενων με το αντικείμενο της.

Είμαι ευγνώμων στον Καθ. Γιάννη Βασιλείου, που ήταν ο επιβλέπων της διατριβής, αρχικά για την εμπιστοσύνη που μου έδειξε και στην συνέχεια για την καθοδήγηση που μου παρείχε όλα αυτά τα χρόνια. Ταυτόχρονα, θα ήθελα να εκφράσω τις θερμότερες ευχαριστίες μου στον Καθ. Τίμο Σελλή για την καθοδήγηση και την υποστήριξή του. Παράλληλα, θα ήθελα να ευχαριστήσω τον Καθ. Κώστα Κοντογιάννη και τον Καθ. Στέφανο Κόλλια για τα σχόλια και τις προτάσεις τους σε θέματα μελλοντικής έρευνας στην ενδιάμεση κρίση.

Επιπλέον, θέλω να ευχαριστήσω ιδιαιτέρως για την εποικοδομητική συνεργασία που είχαμε όλα αυτά τα χρόνια, τον Αν. Καθ. Ιωάννη Αναγνωστόπουλο. Οι συμβουλές και η καθοδήγησή του με βοήθησαν να βελτιώσω την ποιότητα της διατριβής. Επίσης, θέλω να ευχαριστήσω όλους όσους συνεργάστηκα, κατά περιόδους, σε διαφορετικές φάσης της διατριβής μου, και ιδιαίτερα τον Δρ. Γιώργο Παπαστεφανάτο και τον Δρ. Γιώργο Γιαννόπουλο, επιστημονικούς συνεργάτες στο Ινστιτούτο Πληροφοριακών Συστημάτων του Ερευνητικού Κέντρου 'Αθηνά'. Η συγκεκριμένη διατριβή θα ήταν εξαιρετικά δύσκολο να έρθει σε πέρας χωρίς τη βοήθεια, τις ιδέες τους και τα πολύτιμα σχόλιά τους. Επίσης, θέλω να ευχαριστήσω τον Δρ. Γιώργο Κούζα για την ενδιαφέρουσα οπτική του και τις παρατηρήσεις του.

Επιπρόσθετα, οφείλω να ευχαριστήσω όλα τα μέλη του Εργαστηρίου Συστημάτων Βάσεων Ενώσεων και Δεδομένων του ΕΜΠ και του Ινστιτούτου Πληροφοριακών Συστημάτων του Ερευνητικού Κέντρου 'Αθηνά' με τα οποία είχα τη χαρά να συνεργαστώ όλα αυτά τα χρόνια. Παράλληλα θέλω να ευχαριστήσω τον Καθ. Παναγιώτη Τσανάκα και τον Αν. Καθ. Γιώργο Στάμου που με προθυμία δέχτηκαν να συμμετέχουν στη διαδικασία κρίσης της διατριβής, διατελώντας μέλη της επταμελούς εξεταστικής επιτροπής.

Τελευταίο, αλλά εξίσου σημαντικό, οφείλω ένα μεγάλο ευχαριστώ αναγνωρίζοντας την στήριξη, την ενθάρρυνση και κυρίως την υπομονή της συζύγου μου.

Στην Σωτηρία και την Ευαγγελία
Για το παιχνίδι που τους στέρησα

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Περίληψη

Στην εποχή της κοινωνίας της πληροφορίας ο κλάδος της Νομικής Πληροφορικής έχει να αντιμετωπίσει σημαντικές προκλήσεις εξαιτίας του όγκου και της πολυπλοκότητας των νομικών δεδομένων. Σε αυτό το πλαίσιο, ζητήματα διαχείρισης και διάχυσης της νομικής πληροφορίας, μοντέλα διαχείρισης της νομικής πολυπλοκότητας, τεχνικές διευκόλυνσης των χρηστών στην αναζήτηση νομικών πληροφοριών και μέθοδοι ενθάρρυνσης της συμμετοχής των πολιτών στο σχεδιασμό των ρυθμίσεων, αποτελούν ανοιχτά ερευνητικά ζητήματα.

Προς την κατεύθυνση αυτή, η παρούσα διατριβή μελετά και προτείνει μεθόδους: α) διαχείρισης της νομικής πληροφορίας με σημασιολογικά πρότυπα, β) μοντελοποίησης του δικαίου σε μορφή σύνθετου δικτύου, γ) διαφοροποιημένης ανάκτησης νομικής πληροφορίας και δ) διαφοροποιημένης ανάκτησης καταχωρήσεων σε κείμενα διαβουλεύσεων και κοινωνικά δίκτυα.

Στο πλαίσιο της διατριβής προτείνουμε μια πρότυπη μέθοδο για την αυτόματη μοντελοποίηση και σημασιολογική αναπαράσταση νομικών πηγών, μέσω της δημιουργίας μιας γλώσσας συγκεκριμένου τομέα για τις νομικές πηγές. Δεδομένου ότι τα νομικά έγγραφα διαχέονται σε μη μηχαναγνώσιμες μορφές, είναι απαραίτητος ο αυτόματος μετασχηματισμός τους σε μορφή κατάλληλη για τη μοντελοποίηση νομικών πηγών, με σκοπό την δομική και σημασιολογική αναπαράστασή τους, τη διασύνδεσή τους βάσει νομικών παραπομπών και την ταξινόμησή τους. Τα παραπάνω έχουν υλοποιηθεί σε μια πλατφόρμα διαχείρισης νομικής πληροφορίας, η οποία αξιοποιεί την σημασιολογική αναπαράσταση των νομικών πηγών, προσφέροντας, μεταξύ άλλων, εξελιγμένα αποτελέσματα αναζήτησης. Η προτεινόμενη αρχιτεκτονική αξιολογήθηκε σε πραγματικό περιβάλλον παραγωγής, του δημοσίου τομέα, παρέχοντας στο ευρύ κοινό σημασιολογική πρόσβαση στην ελληνική φορολογική νομοθεσία.

Ταυτόχρονα, προτείνουμε ένα μοντέλο αναπαράστασης του γραπτού δικαίου σε μορφή σύνθετου δικτύου. Εφαρμόζουμε το μοντέλο στο σύνολο του γραπτού δίκαιου της Ε.Ε. και εξετάζουμε την δομή και την τοπολογία του, προσπαθώντας να εντοπίσουμε οργανωτικές αρχές του γραπτού δικαίου. Η εμπειρική μας ανάλυση αναδεικνύει σε μακροσκοπικό επίπεδο αφανείς οργανωτικές αρχές του σώματος του δικαίου και παρέχει ερμηνεία για την επίδραση της δομής του δικτύου σε μεμονωμένες νομικές πηγές και σε μικροσκοπικό επίπεδο επιτρέπει την ποσοτικοποίηση της σχετικής σημασίας μιας νομικής πηγής μέσα σε ένα σώμα κειμένων. Προκύπτει, μεταξύ άλλων, ότι οι νομικές πηγές έχουν έντονη τάση να συνδέονται με νομικές πηγές του ιδίου τύπου, σχηματίζοντας ομάδες του ίδιου τύπου/τομέα. Η επικοινωνία μεταξύ των πολύ συσσωρευμένων περιοχών αραιά συνδεδεμένων κόμβων διατηρείται από μερικούς κόμβους, καθώς το δίκτυο είναι επίσης εξαιρετικά ετερογενές σε σχέση με τον αριθμό των

συνδέσεων των νομικών πηγών και συγκεκριμένα είναι ένα δίκτυο νόμου δύναμης μικρού κόσμου (power law small-world network). Η προέλευση αυτής της ετερογένειας, μπορεί να εξηγηθεί από τη διαδικασία της επιλεκτικής προσκόλλησης, η οποία ενισχύει τη δημοτικότητα των πηγών υψηλής κατάταξης. Ταυτόχρονα, αξιολογούμε την χρονική εξέλιξη καθώς και την ανθεκτικότητά του σε περίπτωση μεταβολών. Η πρόταση μας παρέχει μια πρώτη προσέγγιση για την βελτίωση της αποτελεσματικότητας του νομικού συστήματος, ενώ παράλληλα νέες ερευνητικές κατευθύνσεις είναι δυνατό να προκύψουν μέσω αυτής.

Επιπρόσθετα, στην παρούσα διατριβή, εξετάζονται θέματα μεγιστοποίησης της νομικής ποικιλομορφίας των αποτελεσμάτων αναζήτησης, με στόχο την διευκόλυνση των χρηστών κατά την αναζήτηση χρήσιμης πληροφορίας σε ένα τεράστιο όγκο νομικών δεδομένων. Για παράδειγμα, ένας δικηγόρος που προετοιμάζει τα επιχειρήματα του για δεδομένη υπόθεση θα διευκολυνθεί περισσότερο από μια λίστα αποφάσεων που περιέχει αποφάσεις από διαφορετικούς κλάδους, διαφορετικά δικαστήρια, σε διαφορετικές εποχές, σε σχέση με μια λίστα ομοιογενών αποφάσεων με παρόμοια χαρακτηριστικά. Συγκεκριμένα, προσαρμόζουμε αλγορίθμους που έχουν προταθεί στη βιβλιογραφία για την κάλυψη ετερογενών αναγκών, όπως η δημιουργία περιλήψεων κειμένων, η διαφοροποιημένη κατάταξη σε γράφους και η διαφοροποίηση αποτελεσμάτων αναζήτησης. Ταυτόχρονα, εξετάζουμε την συνεισφορά εξειδικευμένων κριτηρίων διαφοροποίησης νομικών πηγών, τα οποία και ενσωματώνουμε στους αλγορίθμους. Πραγματοποιούμε εκτενή πειραματική αξιολόγηση των μεθόδων και κριτηρίων διαφοροποίησης σε ποικίλες περιπτώσεις, με πραγματικές συλλογές νομικών εγγράφων, από διαφορετικά νομικά συστήματα, χρησιμοποιώντας διεθνώς αποδεκτές μετρικές και αντικειμενική μεθοδολογία επισημείωσης του συνόλου δεδομένων, παρέχοντας όρια εξισορρόπησης μεταξύ της σχετικότητας και της ποικιλομορφίας του συνόλου αποτελεσμάτων.

Παράλληλα, με βάση την συμμετοχή των πολιτών στο σχεδιασμό των νομοθετικών ρυθμίσεων μέσω της διαδικασίας διαβούλευσης, αλλά και της ευρείας εξάπλωσης των κοινωνικών δικτύων, εξετάζουμε τη διαφοροποιημένη ανάκτηση καταχωρήσεων χρηστών σε κείμενα διαβουλεύσεων και σε κοινωνικά δίκτυα. Στην κατεύθυνση αυτή, ορίζουμε εξειδικευμένα κριτήρια διαφοροποίησης που λαμβάνουν υπόψη τα χαρακτηριστικά των καταχωρήσεων και του κοινωνικού δικτύου, τα οποία και εισάγουμε σε ευρεστικούς αλγόριθμους διαφοροποίησης, με στόχο την ανάκτηση συνόλου ετερογενών/ποικιλόμορφων καταχωρήσεων. Για τις ανάγκες της πειραματικής αξιολόγησης των μεθόδων/κριτηρίων διαφοροποίησης, που πραγματοποιήθηκε με βάση δημοσίως διαθέσιμα πραγματικά σύνολα δεδομένων, επεκτείναμε μετρικές αξιολόγησης για την αποτίμηση της ποικιλομορφίας των καταχωρήσεων.

Λέξεις Κλειδιά: Νομική Πληροφορική, Διαφοροποίηση, Ανάλυση δικτύων.

Abstract

Information society poses new threats to the legal informatics discipline, mainly due to the volume and complexity of legal data. In this context, legal information management and dissemination, legal complexity, techniques facilitating users in seeking legal information, and methods to encourage citizens' participation in regulatory planning activities are challenging research issues to be addressed.

This doctoral thesis reports upon studies for a) legal sources management with semantic standards; b) modeling civil law as a complex network, c) application of diversification methods for legal information retrieval, and d) application of diversification methods for public consultation texts and social networks.

We present a novel methodology that acquires a semantic representation of legislation, from unstructured formats, by expressing legal documents structure in the form of a set of syntactic rules, i.e., a domain-specific language for legal documents. Since legal documents are usually disseminated in unstructured formats, it is advisable to transform them to another format, suitable for modelling legal sources, capturing the internal organization of the textual structure and the legal semantics, interlinking them based on discovered legal references and classifying them. The above has been integrated on legal document management platform aiming to improve access to legal sources by offering advanced modelling, managing and mining functions. The platform has been successfully deployed in a public sector operated production environment, providing citizens semantic access to Greek tax law.

We also propose a novel approach to model civil law collections as a complex network. We applied our approach on the European Union legislation corpus and identified otherwise, hidden organizing principles of the legislation corpus, interpreted the influence of the network structure to individual legal sources and quantified the relative importance of a legal source within the legislation corpus. Among others, legal sources have a strong tendency to connect with legal documents of the same type, forming clusters of the same sector. Communication between highly clustered areas of sparsely connected nodes is maintained by a few hubs, since the Legislation Network is also highly heterogeneous with respect to the number of edges incident on a node and in particular it is a *small world power law network*. The origin of this heterogeneity may be derived by the preferential attachment process, which amplifies the popularity of highly ranked sources. Further, we studied the temporal evolution of the legislation corpus and evaluated its tolerance to

errors, by performing a resilience test. Our approach aims to improve the efficiency of the legal system and future research directions can be built on our findings.

Additionally, we address diversification of results in legal search as a means of assisting user's searching for useful information in a huge amount of legal data. For example, a lawyer preparing his/her arguments for a given case will find more informative and helpful a diverse result, i.e., a result containing several claims, varying in the type of court and other characteristics, than a set of homogeneous results that contain only relevant cases with similar features. We adopt several state of the art methods from the web search, network analysis and text summarization domains. We also look at the contribution of legal sources diversification criteria, which we also incorporate into the algorithms. We provide an exhaustive evaluation of the methods and criteria in a variety of settings, using real collections of legal documents, from different legal systems, that we objectively annotated with relevance judgments for this purpose, using widely accepted metrics, offering balance boundaries between reinforcing relevant documents or sampling the information space around the legal query

Also, taking into consideration citizen's involvement in regulations through public consultation, as well as the widespread use of social networks, we address result diversification on user comments/microblog post. Towards this direction, we define comment and microblog posts-specific diversification criteria and apply them on heuristic diversification algorithms. We perform an experimental analysis showing that the diversity criteria we introduce result in distinctively diverse subsets of user's posts.

Keywords: Legal Informatics, Diversification, Network analysis

Κεφάλαιο 1

Εισαγωγή

Ο κλάδος της νομικής Πληροφορικής μελετά τις μεθόδους πρόσβασης και επεξεργασίας της νομικής πληροφορίας, με στόχο την ορθολογικότερη και αποτελεσματικότερη οργάνωση και λειτουργία του νομικού συστήματος και της νομικής επιστήμης. Παρότι ο κλάδος δεν είναι καινούργιος [108], καθώς η νομική επιστήμη αποτέλεσε έναν από τους πρώτους τομείς εφαρμογής της επιστήμης των υπολογιστών, η σημασία του σήμερα είναι ιδιαίτερα σημαντική, δεδομένων των πρόσφατων τεχνολογικών εξελίξεων και της έλευσης της κοινωνίας της πληροφορίας [151].

Ανάμεσα στις προκλήσεις που έχει να αντιμετωπίσει ο κλάδος της νομικής πληροφορικής συγκαταλέγονται ζητήματα διαχείρισης και διάχυσης της νομικής πληροφορίας. Παρότι η δημοσίευση δεδομένων της δημόσιας διοίκησης στο διαδίκτυο είναι μια πρακτική που ακολουθούν οι κυβερνήσεις προκειμένου να ενισχύσουν την διαφάνεια και την αξιοπιστία και να εξυπηρετήσουν καλύτερα τον πολίτη, ενθαρρύνοντας ταυτόχρονα την δημόσια και εμπορική χρήση αυτών των δεδομένων, στην πράξη οι νομικές πηγές διαχέονται κυρίως σε μη μηχαναγνώσιμες μορφές [120]. Η πρακτική αυτή καθιστά αδύνατη την αυτόματη επαναχρησιμοποίηση του περιεχομένου και την διασύνδεση με άλλα αποθετήρια και δυσχεραίνει τους τελικούς χρήστες να εντοπίσουν χρήσιμες νομικές πληροφορίες.

Ταυτόχρονα, το γραπτό δίκαιο εξελίσσεται με την πάροδο του χρόνου, καθώς καινούργια νομικά έγγραφα συνεχώς δημιουργούνται και υφιστάμενα τροποποιούνται ή ακυρώνονται. Η εξέλιξη αυτή, σε συνδυασμό με την αλληλεξάρτηση των νομικών πηγών και τον αριθμό των αναφορών που απαιτούνται για την ερμηνεία τους, αυξάνει το βαθμό δυσκολίας εύρεσης, κατανόησης, αλλά και συστηματικής ερμηνείας του ρυθμιστικού πλαισίου τόσο για τους πολίτες, όσο και για τους ειδικούς του χώρου.

Επιπρόσθετα, οι πρωτοβουλίες ανοιχτών δεδομένων συνετέλεσαν σε τεράστια αύξηση του αριθμού των νομικών πηγών που είναι ελεύθερα διαθέσιμες. Η αύξηση όμως αυτή, δε συνοδεύτηκε από ανάλογη παροχή, εξειδικευμένων στο νομικό γίγνεσθαι, τεχνικών εντοπισμού χρήσιμης πληροφορίας, με στόχο τη διευκόλυνση των χρηστών στην εξεύρεση νομικής πληροφορίας.

Τέλος, η δημόσια συμμετοχή στη λήψη αποφάσεων μέσω και της διαδικασίας διαβούλευσης μέσω διαδικτύου εισάγει θέματα διαχείρισης και επεξεργασίας δεδομένων που οφείλουν να

εξεταστούν.

Στο πλαίσιο της παρούσας διατριβής προτείνονται μέθοδοι για τα εξής προβλήματα:

1. Μοντελοποίηση και Διαχείριση Νομοθεσίας. Προτείνουμε και υλοποιούμε την αρχιτεκτονική μιας πλατφόρμας διαχείρισης νομικών πηγών, που παρέχει προηγμένες υπηρεσίες μοντελοποίησης, διαχείρισης και ανάκτησης νομικής πληροφορίας.

2. Νομική Πολυπλοκότητα. Προτείνουμε ένα μοντέλο αναπαράστασης του γραπτού δικαίου σε μορφή σύνθετου δικτύου και μελετάμε την εφαρμογή του στο δίκαιο της Ε.Ε.

3. Ανάκτηση Νομικής Πληροφορίας. Προτείνουμε και αξιολογούμε μεθόδους για την μεγιστοποίηση της νομικής ποικιλομορφίας των αποτελεσμάτων αναζήτησης.

4. Κοινωνικές Διαστάσεις. Εξετάζουμε θέματα διαφοροποιημένης ανάκτησης καταχωρήσεων χρηστών σε κείμενα διαβουλεύσεων και σε κοινωνικά δίκτυα και προτείνουμε/αξιολογούμε μεθόδους.

1.1 Προβλήματα και Προκλήσεις

1.1.1 Μοντελοποίηση και Διαχείριση Νομοθεσίας

Για την επίτευξη των στόχων του σημασιολογικού ιστού στον νομικό τομέα [16], απαραίτητη προϋπόθεση αποτελεί η δόμηση και διάχυση των νομικών πηγών σε μια τυποποιημένη μορφή. Όπως προκύπτει από σχετική μελέτη [120], σε παγκόσμια βάση, για το έτος 2016, μόλις το 26% των κοινοβουλίων χρησιμοποιεί κάποιο πρότυπο του σημασιολογικού ιστού ως μορφή διανομής νομικών πηγών. Αντίθετα, οι νομικές πηγές προσφέρονται στους χρήστες σε αδόμητη μορφή, προσανατολισμένη στην παρουσίαση, καθιστώντας πρακτικά αδύνατη την επαναχρησιμοποίηση του περιεχομένου, τη διασύνδεση με άλλα αποθετήρια στον σημασιολογικό ιστό και δυσχεραίνοντας τελικά τους τελικούς χρήστες στην προσπάθεια τους να εντοπίσουν νομικές πληροφορίες που θα ικανοποιήσουν την πληροφοριακή τους ανάγκη.

Ένα μοντέλο αναπαράστασης νομικής γνώσης είναι απαραίτητο για τη δόμηση των νομικών εγγράφων, την αναπαράσταση των μεταδεδομένων και των συνδέσεων, και τη μορφοποίηση νομικών εγγράφων. Διάφορα μοντέλα έχουν προταθεί ως αποτέλεσμα εθνικών ή διεθνών προσπαθειών [22, 93, 14]. Καθένα από αυτά παρουσιάζει συγκεκριμένα πλεονεκτήματα και αδυναμίες και επιπρόσθετα κανένα από αυτά δεν έχει επιχειρηθεί στο παρελθόν να επεκταθεί καταλλήλως, προκειμένου να έχει εφαρμογή στις Ελληνικές νομικές πηγές.

Η ύπαρξη ενός μοντέλου αναπαράστασης νομικής γνώσης αποτελεί το σημείο αφετηρίας. Για να έχει νόημα στην πράξη ένα τέτοιο μοντέλο θα πρέπει να τύχει και ανάλογης εφαρμογής, δηλαδή να χρησιμοποιηθεί έμπρακτα στις νομικές πηγές. Προς αυτή την κατεύθυνση, έχει προταθεί εξειδικευμένο λογισμικό επεξεργασίας [109, 1], όπου ο συντάκτης υποχρεούται να παρέχει πληροφορίες σχετικά με το κείμενο κατά την συγγραφή. Η τεχνική αυτή, μολονότι επιφέρει βέλτιστα αποτελέσματα, απαιτεί εκτεταμένη εργασία, εξειδικευμένου προσωπικού, ιδίως εάν αναλογιστούμε τον όγκο του υφιστάμενου νομικού έργου (legacy legal data).

Στην παρούσα διατριβή, προτείνουμε μια εναλλακτική προσέγγιση, την αυτόματη κατασκευή μιας μηχανικά αναγνώσιμης μορφής αναπαράστασης των νομικών πηγών, η οποία έχει εφαρμογή και στις υφιστάμενες νομικές πηγές. Η μέθοδος μας βασίζεται στην παρατήρηση ότι οι νομικές πηγές ακολουθούν 'αυστηρή' γλώσσα σύνταξης και καθορισμένη λεκτική δομή διατύπωσης/διάρθρωσης. Εκφράσαμε την δομή των νομικών πηγών με την μορφή ενός συνόλου συντακτικών κανόνων και αναπτύξαμε μια γλώσσα ειδικού σκοπού για την δημιουργία ενός συντακτικού αναλυτή νομικών πηγών.

Η ύπαρξη των νομικών πηγών σε μια τυποποιημένη μορφή, με βάση διεθνώς αποδεκτά πρότυπα του σημασιολογικού ιστού [17], αποτελεί μέρος ενός ευρύτερου προβλήματος, αυτού της διαχείρισης και διάχυσης της νομικής πληροφορίας. Η αρχική ερευνητική κατεύθυνση για την δημιουργία έμπειρων νομικών συστημάτων πλέον εστιάζει στον τομέα των συστημάτων διαχείρισης νομικών πηγών και νομικής γνώσης. Οι υπάρχουσες όμως προτάσεις στην βιβλιογραφία [21, 62, 50], εστιάζουν σε ορισμένα από τα χαρακτηριστικά του προβλήματος, χωρίς να παρέχουν μια συνολική αντιμετώπιση των προκλήσεων που πρέπει να αντιμετωπιστούν. Συγκεκριμένα η πλειοψηφία των προτεινόμενων συστημάτων αγνοεί την διαδικασία μοντελοποίησης των νομικών πηγών, θεωρώντας ότι αυτά βρίσκονται διαθέσιμα σε μια ιδανική μορφή. Επιπρόσθετα, κανένα από τα προτεινόμενα συστήματα δεν παρέχει προηγμένες υπηρεσίες αναζήτησης νομικής πληροφορίας βασισμένες σε τεχνικές δομημένης ανάκτησης και διαφοροποιημένης ανάκτησης νομικής πληροφορίας. Επίσης, κανένα από τα προτεινόμενα συστήματα δεν ενσωματώνει την οπτική των χρηστών ως μέσο παρουσίασης/ εξερεύνησης της νομικής πληροφορίας.

Στην παρούσα διατριβή προτείνουμε μια αρχιτεκτονική για την διαχείριση νομικών πηγών, που αυτόματα, συλλέγονται, μοντελοποιούνται, διασυνδέονται και ταξινομούνται, παρέχοντας στους χρήστες, μεταξύ άλλων, στοχευμένα και εμπλουτισμένα αποτελέσματα αναζήτησης

1.1.2 Νομική Πολυπλοκότητα

Οι κανονισμοί και οι οδηγίες, που προέρχονται από τις αρχές σε Ευρωπαϊκό, Εθνικό και Τοπικό επίπεδο και επηρεάζουν τις διάφορες πτυχές των καθημερινών δραστηριοτήτων, παρά τις προσπάθειες εναρμόνισης και απορρύθμισης, αυξάνονται συνεχώς. Ταυτόχρονα, υφιστάμενοι κανονισμοί και οδηγίες τροποποιούνται ή ακυρώνονται. Οι κανονισμοί αυτοί εντάσσονται σε ένα ευρύτερο σύστημα ρύθμισης με βάση το οποίο και ερμηνεύονται. Το πλέγμα αυτό, αλληλο-συσχετίσεων και αλληλο-συνδέσεων, επηρεάζει τόσο τους πολίτες, που καλούνται να σεβαστούν και να εφαρμόσουν το ισχύον δίκαιο, όσο και τα νομοθετικά και εκτελεστικά όργανα που καλούνται να συντάξουν νέες νομοθετικές ρυθμίσεις και να τροποποιήσουν υφιστάμενες έχοντας πλήρη γνώση των συσχετίσεων μεταξύ των πηγών δικαίου.

Οι μέχρι τώρα προταθείσες στην βιβλιογραφία προσπάθειες, από την πλευρά της επιστήμης των υπολογιστών, συνίσταται στην ανάλυση του δικτύου παραπομπών μεταξύ δικαστικών αποφάσεων [47, 48, 135, 91], με έμφαση στο Αμερικανικό νομικό σύστημα, με κύριο στόχο την ανακάλυψη δεδικασμένων. Το δεδικασμένο αποτελεί νομικό κανόνα, που προέρχεται από το Αγγλικό δίκαιο, και ενθαρρύνει τους δικαστές να ακολουθήσουν προηγούμενη δικαστική

απόφαση, η οποία αποτελεί 'νόμο' για την έκδοση απόφασης σε παρόμοια υπόθεση. Στην ελληνική έννομη τάξη, όπως και στις υπόλοιπες έννομες τάξεις της Ευρώπης, που έλκουν το νομικό τους σύστημα από το Ρωμαϊκό Δίκαιο, οι πηγές των κανόνων δικαίου που διέπουν την έννομη τάξη είναι πολύ συγκεκριμένες. Αυτά τα συστήματα δικαίου αποκαλούνται civil law και σε αντίθεση με το Αγγλοσαξονικό δίκαιο (common law) οι αποφάσεις των δικαστηρίων δεν αποτελούν πηγή του Δικαίου. Τα δικαστήρια ερμηνεύουν του ισχύοντες νόμους, οι δε αποφάσεις τους αποτελούν πηγή ερμηνείας του Δικαίου.

Οι προηγούμενες μελέτες έχουν κατά κύριο λόγο εφαρμοστεί σε νομικά συστήματα συστήματα common law και δεν λαμβάνουν υπόψη τους τα ιδιαίτερα χαρακτηριστικά των συστημάτων civil law. Στην παρούσα διατριβή, προτείνουμε ένα μοντέλο αναπαράστασης του δικαίου, το οποίο παρέχει μια συστηματική εναλλακτική δομή για την περιγραφή ενός φυσικά εξελισσόμενου κανονιστικού πλαισίου, με στόχο τη δημιουργία ενιαίου υποβάθρου για την μεταφορά από την νομική στην μηχανική. Το Δίκτυο Νομοθεσίας είναι ένα σύνθετο δίκτυο που περικλείει την ιεραρχία μεταξύ των πηγών του δικαίου και μπορεί να αναπαραστήσει συσχετίσεις διαφόρων κατηγοριών μεταξύ των νομικών πηγών, παράλληλα με την χρονική εξέλιξή τους.

Εφαρμόζουμε το μοντέλο στο σύνολο του γραπτού δίκαιου της Ε.Ε. και εξετάζουμε την δομή και την τοπολογία του προσπαθώντας να εντοπίσουμε οργανωτικές αρχές του γραπτού δικαίου. Στην συνέχεια, μελετάμε πώς το δίκαιο εξελίσσεται με την πάροδο του χρόνου, καθώς καινούργιες ρυθμίσεις τίθενται σε ισχύ ενώ άλλες τροποποιούνται ή καταργούνται. Τέλος, αξιολογούμε την ανοχή του δικτύου σε αλλαγές, πραγματοποιώντας μια δοκιμή ανθεκτικότητας. Η εργασία μας παρέχει μια πρώτη προσέγγιση για την παροχή ενός μοντέλου για να εξηγήσουμε καλύτερα τη δομή και την εξέλιξη του δικαίου. Νέες ερευνητικές κατευθύνσεις για τη βελτίωση της αποτελεσματικότητας του νομικού συστήματος είναι δυνατό να προκύψουν μέσω αυτής.

1.1.3 Ανάκτηση Νομικής Πληροφορίας

Στην σημερινή εποχή, ως συνέπεια των πρωτοβουλιών ανοιχτών δεδομένων, παρατηρείται μια τεράστια αύξηση στον αριθμό των συνόλων νομικών δεδομένων που είναι ελεύθερα διαθέσιμα. Νομικά δεδομένα και κείμενα που ήταν προηγουμένως διαθέσιμα μόνο σε ένα εξειδικευμένο κοινό είναι τώρα ελεύθερα διαθέσιμα στο διαδίκτυο. Κυβερνητικές ή ιδιωτικές πύλες παρέχουν πρόσβαση σε ολοένα και αυξανόμενο αριθμό κανονισμών, δικαστικών υποθέσεων ή διοικητικών αποφάσεων, προσφέροντας συνήθως υπηρεσίες θεματικής πλοήγησης και αναζήτησης με λέξεις κλειδιά.

Η τεράστια αύξηση στον αριθμό των συνόλων νομικών πηγών που είναι ελεύθερα διαθέσιμες στο διαδίκτυο εισάγει ταυτόχρονα θέματα διαχείρισης και επεξεργασίας δεδομένων, λόγω του μεγάλου όγκου πληροφορίας προς επεξεργασία (information overload). Σε ένα τέτοιο σενάριο, ο εντοπισμός χρήσιμης πληροφορίας σε ένα τεράστιο όγκο δεδομένων αποτελεί μια εξαιρετικά δύσκολη διεργασία.

Οι τεχνικές ανάκτησης νομικής πληροφορίας που έχουν προταθεί στην βιβλιογραφία, συ-

νίσταται κυρίως σε εξωτερικές πηγές γνώσης πχ θησαυροί, οντολογίες. Θέματα μεγιστοποίησης της νομικής ποικιλομορφίας των αποτελεσμάτων αναζήτησης, με στόχο την διευκόλυνση των χρηστών κατά την αναζήτηση πληροφορίας, δεν έχουν επί του παρόντος διερευνηθεί.

Ας σκεφτούμε, για παράδειγμα, την περίπτωση ενός δικηγόρου ο οποίος προετοιμάζει τα επιχειρήματα του για δεδομένη υπόθεση και υποβάλλει ένα ερώτημα χρήστη για την ανάκτηση σχετικών πληροφοριών. Θα πρέπει επαναληπτικά να περιηγηθεί σε μια τεράστια λίστα αποφάσεων, επιλέγοντας κάθε φορά μέσω της εμπειρίας του τα σχετικά έγγραφα, προκειμένου να αποκτήσει ένα ευρύ, σε πλάτος και σε βάθος, πλαίσιο κατανόησης. Ένα διαφοροποιημένο αποτέλεσμα, δηλαδή, ένα αποτέλεσμα που περιέχει αρκετές αξιώσεις, που καλύπτει ένα ετερογενές φάσμα δυνατών νομικών ερμηνειών, σε διαφορετικές εποχές, είναι διαισθητικά πιο κατατοπιστικό από ένα ομοιογενές σύνολο αποτελεσμάτων συναφών αποφάσεων με παρόμοια χαρακτηριστικά.

Οι τεχνικές διαφοροποίησης αποτελεσμάτων αναζήτησης έχουν προταθεί στην βιβλιογραφία, για τη βελτίωση της ικανοποίησης του χρήστη, μέσω της αύξησης της ποικιλίας των πληροφοριών που λαμβάνει ο χρήστης από τα αποτελέσματα αναζήτησης.

Στην παρούσα διατριβή, προτείνουμε και εξετάζουμε την συνεισφορά εξειδικευμένων κριτηρίων διαφοροποίησης νομικών πηγών, λαμβάνοντας υπόψιν τα χαρακτηριστικά τους. Τα κριτήρια αυτά, τα ενσωματώνουμε σε αλγορίθμους που έχουν προταθεί στην βιβλιογραφία για την κάλυψη ετερογενών αναγκών, όπως η διαφοροποίηση αποτελεσμάτων αναζήτησης, η διαφοροποιημένη κατάταξη σε γράφους και η δημιουργία περιλήψεων κειμένων. Τους αλγορίθμους, που ανήκουν στις προαναφερθείσες κατηγορίες, τους προσαρμόζουμε στο συγκεκριμένο πρόβλημα και στα κριτήρια που προτείνουμε και μελετάμε διεξοδικά την απόδοσή τους σε πραγματικά δεδομένα από διαφορετικά νομικά συστήματα.

1.1.4 Κοινωνικές Διαστάσεις

Στο πλαίσιο πρωτοβουλιών για την Ανοικτή Δημόσια Διοίκηση και Διακυβέρνηση, έχει διεθνώς καθιερωθεί η δημόσια συμμετοχή στη λήψη αποφάσεων. Αρκετά διαδεδομένη μορφή αυτής, αποτελεί η διαδικασία διαβούλευσης μέσω του διαδικτύου για την αξιοποίηση των σχολίων και προτάσεων των πολιτών. Αυτό το είδος της διαβούλευσης επιτρέπει στους πολίτες να εκφράσουν απόψεις, προτάσεις, επιχειρήματα, διαφωνίες που σχετίζονται με την εν εξελίξει νομοθεσία.

Η πρακτική αυτή προάγει την διαφάνεια και ενθαρρύνει την συμμετοχή των πολιτών στο σχεδιασμό των ρυθμίσεων, αλλά ταυτόχρονα, εισάγει θέματα διαχείρισης και επεξεργασίας δεδομένων, λόγω του μεγάλου και ετερογενούς όγκου πληροφορίας προς επεξεργασία. Πολλές φορές τα κείμενα διαβούλευσης, ενδέχεται να συγκεντρώνουν εκατοντάδες σχόλια χρηστών, γεγονός που καθιστά την επισκόπηση του συνόλου των σχολίων από τους χρήστες ιδιαίτερα χρονοβόρα. Αρκετές φορές, επίσης, το περιεχόμενο του κειμένου από μόνο του δεν είναι αρκετό για να σχηματίσει ο χρήστης μία πλήρη εικόνα πάνω στα θέματα τα οποία πραγματεύεται. Η κοινή γνώμη είναι ένας σημαντικός παράγοντας που συμπληρώνει το αντικείμενο της διαβούλευσης και αντιπροσωπεύει τη 'σοφία του πλήθους'. Σε αυτήν την περίπτωση, ο χρήστης

χρειάζεται να μπορεί να δει ένα μικρό και όσο το δυνατόν πιο ετερογενές υποσύνολο σχολίων, το οποίο αντιπροσωπεύει διάφορες εκφάνσεις των θεμάτων και γνώμες/συναισθήματα των χρηστών.

Ταυτόχρονα, τα κοινωνικά δίκτυα έχουν γίνει πρόσφατα ένα παγκόσμια δημοφιλές μέσο για την παρακολούθηση και τη διάδοση τάσεων, ειδήσεων και ιδεών, καθώς και προσωπικών απόψεων και συναισθημάτων σε ένα ευρύ φάσμα θεμάτων. Εκτός από την ανάγνωση περιεχομένου στις δικές τους ροές, οι χρήστες των κοινωνικών δικτύων συχνά αναζητούν στο δημοσιευμένο περιεχόμενο, ο αριθμός δε των ερωτημάτων που υποβάλλονται στο Twitter, ημερησίως, ξεπερνά τα δύο δισεκατομμύρια [24]. Σε αυτό το σενάριο, ο τεράστιος όγκος των καταχωρήσεων καθιστά αδύνατη την επισκόπηση του συνόλου των καταχωρήσεων από τους χρήστες. Επομένως, για να σχηματίσει ο χρήστης μία πλήρη εικόνα πάνω στα θέματα ενδιαφέροντος του θα πρέπει να μπορεί να δει ένα μικρό και όσο το δυνατόν πιο ετερογενές υποσύνολο καταχωρήσεων, το οποίο και να αντιπροσωπεύει διάφορες εκφάνσεις, γνώμες και συναισθήματα των χρηστών, σχετικά με την αναζήτησή του.

Σε αντίθεση με τον παραδοσιακό παγκόσμιο ιστό, ο οποίος σε μεγάλο βαθμό οργανώνεται από το περιεχόμενο, τα κοινωνικά δίκτυα ενσωματώνουν τους χρήστες ως κεντρικές οντότητες. Οι χρήστες σε ένα κοινωνικό δίκτυο δημοσιεύουν το δικό τους περιεχόμενο και δημιουργούν συνδέσεις με άλλους χρήστες. Σε ένα τέτοιο σενάριο τεχνικές διαφοροποίησης αποτελεσμάτων αναζήτησης, όπως η διαφοροποίηση βασιζόμενη μόνο στο κειμενικό περιεχόμενο ενός πόρου, οι οποίες λειτουργούν αποτελεσματικά με βάση τα χαρακτηριστικά του παγκοσμίου ιστού, είναι ανεπαρκείς στο σενάριο διαφοροποίησης καταχωρήσεων χρηστών σε κοινωνικά δίκτυα.

Στην κατεύθυνση αυτή, ορίζουμε εξειδικευμένα κριτήρια διαφοροποίησης που λαμβάνουν υπόψη τα χαρακτηριστικά των καταχωρήσεων/σχολίων, σημασιολογικά μεταδεδομένα των καταχωρήσεων/σχολίων, και του κοινωνικού δικτύου, μεταδεδομένα των καταχωρήσεων προερχόμενα από τα χαρακτηριστικά του δικτύου, υποστηρίζοντας ότι η ετερογένεια σε αυτά τα κριτήρια συνεπάγεται και ετερογένεια στα θέματα/έννοιες/απόψεις που περιέχονται στις καταχωρήσεις/σχόλια. Ενσωματώνουμε τα κριτήρια αυτά σε ευριστικούς αλγόριθμους διαφοροποίησης, ώστε να παράγουμε ένα υποσύνολο των αρχικών καταχωρήσεων που περιέχει ετερογενείς καταχωρήσεις. Επιπρόσθετα, επεκτείνουμε υπάρχουσες έννοιες και μετρικές για την αξιολόγηση του αποτελέσματος της διαφοροποίησης των καταχωρήσεων.

1.2 Συνεισφορά

Επιγραμματικά η συνεισφορά της διατριβής συνοψίζεται στα παρακάτω σημεία:

– προτείνουμε μια πρότυπη μέθοδο για την αυτόματη μοντελοποίηση και σημασιολογική αναπαράσταση νομικών πηγών, σε ένα μοντέλο σημασιολογικής αναπαράστασης νομικής πληροφορίας που επεκτείνουμε για την κάλυψη των ιδιαιτεροτήτων των ελληνικών νομικών πηγών. Η μέθοδος μας στηρίζεται στην δημιουργία μιας γλώσσας συγκεκριμένου τομέα για τις νομικές πηγές. Ταυτόχρονα προτείνουμε την αρχιτεκτονική μιας πλατφόρμας διαχείρισης νομικών πηγών, η οποία έχει στόχο την βελτίωση της πρόσβασης σε

αυτές, παρέχοντας προηγμένες υπηρεσίες μοντελοποίησης, διαχείρισης και ανάκτησης νομικής πληροφορίας.

- προτείνουμε ένα μοντέλο αναπαράστασης του γραπτού δικαίου σε μορφή σύνθετου δικτύου και πραγματοποιούμε εμπειρική ανάλυση της δομής και τοπολογίας του δικτύου, που προκύπτει από την εφαρμογή του μοντέλου στο σύνολο του γραπτού δίκαιου της Ε.Ε. Ταυτόχρονα, αξιολογούμε την χρονική εξέλιξη καθώς και την ανθεκτικότητα του σε περίπτωση μεταβολών.
- εξετάζουμε και αξιολογούμε θέματα μεγιστοποίησης της νομικής ποικιλομορφίας των αποτελεσμάτων αναζήτησης, προσαρμόζοντας αλγορίθμους που έχουν προταθεί για την κάλυψη ετερογενών αναγκών, όπως η δημιουργία περιλήψεων κειμένων, η διαφοροποιημένη κατάταξη σε γράφους και η διαφοροποίηση αποτελεσμάτων αναζήτησης, προτείνοντας παράλληλα εξειδικευμένα κριτήρια διαφοροποίησης νομικών πηγών.
- αξιολογούμε την απόδοση ευρετικών αλγορίθμων διαφοροποίησης και εξειδικευμένων κριτηρίων που προτείνουμε, για την διαφοροποιημένη ανάκτηση καταχωρήσεων χρηστών σε κείμενα διαβουλεύσεων και σε κοινωνικά δίκτυα.

1.3 Δομή της Διατριβής

Η παρούσα εργασία χωρίζεται εννοιολογικά σε επτά (7) μέρη. Στο εισαγωγικό μέρος οριοθετείται το αντικείμενο και σκιαγραφούνται οι στόχοι της. Στο Κεφάλαιο 2 δίνεται συνοπτικά το υπόβαθρο στο οποίο κινείται η θεματολογία της διατριβής καθώς και μια επισκόπηση συναφών εργασιών.

Στο Κεφάλαιο 3 προτείνουμε μια πρότυπη μέθοδο για την αυτόματη μοντελοποίηση και σημασιολογική αναπαράσταση νομικών πηγών, μέσω της δημιουργίας μιας γλώσσας συγκεκριμένου τομέα για τις νομικές πηγές. Παράλληλα, προτείνουμε την αρχιτεκτονική μιας πλατφόρμας διαχείρισης νομικών πηγών, η οποία έχει στόχο την βελτίωση της πρόσβασης σε νομικές πηγές παρέχοντας προηγμένες υπηρεσίες μοντελοποίησης, διαχείρισης και ανάκτησης νομικής πληροφορίας. Η προσέγγιση μας έχει προταθεί στις εργασίες [78, 77].

Στο Κεφάλαιο 4 προτείνουμε ένα μοντέλο αναπαράστασης του γραπτού δικαίου σε μορφή σύνθετου δικτύου. Μελετάμε την δομή και την τοπολογία του δικτύου που προκύπτει από την εφαρμογή του μοντέλου στο σύνολο του γραπτού δίκαιου της Ε.Ε., αξιολογούμε την χρονική εξέλιξη και την ανθεκτικότητα του σε περίπτωση μεταβολών. Οι μέθοδοι μας έχουν παρουσιαστεί στις εργασίες [75, 71].

Στο Κεφάλαιο 5 προσαρμόζουμε στο πρόβλημα μεγιστοποίησης της νομικής ποικιλομορφίας των αποτελεσμάτων αναζήτησης και σε εξειδικεύμενα κριτήρια που εισάγουμε, διαδεδομένους αλγόριθμους που έχουν προταθεί για την κάλυψη διαφορετικών αναγκών π.χ., την δημιουργία περιλήψεων, την διαφοροποιημένη κατάταξη σε γράφους, παράλληλα με αλγόριθμους διαφοροποίησης αποτελεσμάτων αναζήτησης. Τέλος, πραγματοποιούμε εκτενή πειραματική αξιολόγηση των προαναφερθέντων μεθόδων και κριτηρίων διαφοροποίησης σε ποικίλες περιπτώσεις. Οι μέθοδοι μας έχουν δημοσιευτεί στις εργασίες [72, 73, 74].

Στο Κεφάλαιο 6 αξιολογούμε μεθόδους και κριτήρια διαφοροποίησης καταχωρήσεων χρηστών σε κείμενα διαβουλεύσεων και σε κοινωνικά δίκτυα. Η προσέγγιση μας έχει δημοσιευτεί στις εργασίες [54, 76].

Τέλος, στο Κεφάλαιο 7 συνοψίζουμε την παρουσίαση της εργασίας και καταγράφουμε πιθανές προεκτάσεις των θεμάτων που μελετήθηκαν.

Κεφάλαιο 2

Εννοιολογικό Υπόβαθρο

Αυτό το κεφάλαιο παραθέτει το βασικό υπόβαθρο που χρειάζεται για την κατανόηση της διατριβής και παρουσιάζει τις σημαντικότερες ερευνητικές εργασίες σε θέματα που άπτονται της παρούσης εργασίας.

2.1 Εισαγωγή

Η συγκεκριμένη ενότητα εισάγει τον αναγνώστη σε βασικά προαπαιτούμενα. Αναλυτικότερα στην υπό-ενότητα 2.1.1 παρουσιάζονται κύρια χαρακτηριστικά της Τοπολογίας Δικτύων και στην υπό-ενότητα 2.1.2 Τεχνικές Διαφοροποίησης.

2.1.1 Τοπολογία Δικτύων

Ένα δίκτυο, που ονομάζεται επίσης και γράφος, είναι ένα σύνολο από στοιχεία, που ονομάζονται κόμβοι ή κορυφές, με συνδέσεις μεταξύ τους, που ονομάζεται ακμές ή τόξα. Πιο τυπικά, ένα γράφημα είναι ένα διατεταγμένο ζεύγος $G = (V, A)$ που αποτελείται από ένα σύνολο V των κόμβων με ένα σύνολο A ακμών, τα οποία αποτελούν υποσύνολα 2-στοιχείων του V. Αυτό το φαινομενικά απλό πλαίσιο μοντελοποίησης μπορεί να γίνει πιο σύνθετο αν κάποιος το επεκτείνει προκειμένου να περιλαμβάνει πρόσθετα επίπεδα λεπτομέρειας. Για παράδειγμα μπορεί να υπάρχουν περισσότερα τους ενός είδη κόμβων/ακμών, οι κόμβοι και οι ακμές θα μπορούσαν να εμφανίζονται και εξαφανίζονται εντός χρονικών περιόδων, οι ακμές θα μπορούσαν να έχουν συγκεκριμένη κατεύθυνση, θα μπορούσαν να υπάρχουν πολλαπλές ακμές μεταξύ του ίδιου ζεύγους κόμβων κ.λπ. Κατά συνέπεια, δίκτυα των οποίων η δομή είναι ακανόνιστη, σύνθετη και δυναμικά εξελισσόμενη στον χρόνο μπορούν να σχηματιστούν και να χρησιμοποιηθούν για να περιγράψουν μια ευρεία ποικιλία υποκείμενων συστημάτων.

Συνήθως, τα δίκτυα αποτελούν μαθηματικές αναπαραστάσεις πολύπλοκων συστημάτων π.χ., εξάπλωση ιού σε κοινωνικό δίκτυο, τα οποία και μας βοηθούν να κατανοήσουμε πληρέστερα. Αναλύοντας τη δομή του δικτύου, μπορεί κανείς να αποκαλύψει σημαντικά στοιχεία για τη συμπεριφορά του π.χ., να προβλέψει πόσο γρήγορα ένας ιός θα εξαπλωθεί [113], ποιοί είναι οι πιο σημαντικοί κόμβοι [12, 82] ή να προβλέψει την ανθεκτικότητα μιας περιοχής του δικτύου [6]. Διάφορα μοντέλα δικτύων έχουν προταθεί στη βιβλιογραφία για να μας βοηθήσουν

να κατανοήσουμε ή ακόμα και να προβλέψουμε τη συμπεριφορά φυσικών ή ανθρωπογενών συστημάτων, π.χ., τα δίκτυα μεταφοράς [131], το διαδίκτυο [80], δίκτυο τροφικών πλεγμάτων [101], το δίκτυο των μεταβολικών οδών [68], τα δίκτυα των νευρώνων του εγκεφάλου [15], τα κοινωνικά δίκτυα [98] και πολλά άλλα.

Η έρευνα έχει δείξει ότι τα πραγματικά δίκτυα δεν προκύπτουν από τυχαίες ενώσεις ομοειδών στοιχείων, αλλά εμφανίζουν γενικές οργανωτικές αρχές [103]. Μεταξύ των μαθηματικών ιδιοτήτων που χαρακτηρίζουν αυτές τις αρχές είναι οι εξής:

– Κατανομή βαθμών κόμβων.

Ορίζουμε βαθμό (degree) ενός κόμβου το σύνολο των ακμών με τις οποίες διασυνδέεται με άλλους κόμβους στο δίκτυο. Ανάλογα με το αν το δίκτυο είναι κατευθυνόμενο ή όχι, ο βαθμός ενός δεδομένου κόμβου μπορεί να διακριθεί σε βαθμό εισερχόμενων και εξερχόμενων ακμών. Το άθροισμα των δύο ισούται σε αυτήν την περίπτωση με το συνολικό βαθμό του κόμβου. Η μελέτη των βαθμών των κόμβων μπορεί να μας δώσει άμεσα πολλές πληροφορίες για συγκεκριμένα ιδιαίτερα στατιστικά χαρακτηριστικά του δικτύου. Για παράδειγμα, ο απλός υπολογισμός του βαθμού όλων των κόμβων ενός δικτύου μας βοηθάει να εντοπίσουμε τους κόμβους που είναι περισσότερο διασυνδεδεμένοι στο δίκτυο, εκείνα δηλαδή τα στοιχεία που αλληλεπιδρούν πιο έντονα με τα υπόλοιπα. Δεδομένου ότι δεν έχουν όλοι οι κόμβοι σε ένα δίκτυο έχουν τον ίδιο αριθμό ακμών, η κατανομή βαθμών $P(k)$ ενός δικτύου ορίζεται ως το κλάσμα των κόμβων στο δίκτυο με βαθμό k. Με άλλα λόγια, η συνάρτηση κατανομής $P(k)$, δίνει την πιθανότητα ένας τυχαίο κόμβος να έχει ακριβώς k ακμές. Σε τυχαία δίκτυα [38], όπου οι ακμές τοποθετούνται τυχαία, η πλειοψηφία των κόμβων έχουν περίπου τον ίδιο βαθμό, κοντά στο μέσο όρο βαθμού k του δικτύου και η κατανομή βαθμών είναι μια κατανομή Poisson.

Σε αντίθεση με τα τυχαία δίκτυα, για ένα μεγάλο αριθμό πραγματικών δικτύων, υπάρχουν λίγοι κόμβοι πολύ μεγάλου βαθμού ενώ η συντριπτική πλειοψηφία των κόμβων έχει μόνο μερικές συνδέσεις με άλλους κόμβους [5]. Η μορφή κατανομής βαθμών αυτών των δικτύων, φαίνεται να αγνοεί το μέγεθος του δικτύου και να μην επιτρέπει τον εύκολο υπολογισμό μιας μέσης τιμής για το βαθμό των κόμβων. Δεδομένου ότι το εύρος των τιμών βαθμών κόμβου ποικίλλει σε πολύ μεγάλο βαθμό, ονομάζουμε τα δίκτυα που ακολουθούν αυτές τις κατανομές βαθμών κόμβων 'δίκτυα ανεξάρτητα κλίμακας' (scale-free networks).

Σε πολλές περιπτώσεις, η κατανομή βαθμών $P(k)$ ακολουθεί το νόμο δύναμης (power-law) [31]. Αυτό σημαίνει ότι το κλάσμα των κόμβων με βαθμό k, $P(k)$ ακολουθεί την $K^{-\gamma}$, για κάποιο γ. Η παράμετρος γ, είναι μία πραγματική σταθερά, γνωστή ως εκθέτης ή παράμετρος μείωσης, που τυπικά κυμαίνεται στα όρια $2 < \gamma < 3$ για πραγματικά δίκτυα, αν και μπορεί να βρίσκεται και έξω από αυτά τα όρια. Οι νόμοι δύναμης είναι κατανομές που περιγράφουν χαρακτηριστικά ανεξάρτητα κλίμακας με την έννοια ότι οι μεταβλητές μεταβάλλονται με τρόπο που δεν επηρεάζεται από την κλίμακα και μπορούν να προκύψουν από συγκεκριμένα δυναμικά συστήματα με κοινό χαρακτηριστικό την εξέλιξη στο χρόνο, σηματοδοτώντας κάποιες σημαντικές συσχετίσεις στο εσωτερικό του συστήματος. Πέρα από την προέλευσή τους, τα δομικά τους χαρακτηριστικά κάνουν τέτοια δίκτυα εξαιρετικά

αποτελεσματικά στη μετάδοση πληροφορίας, την ταχεία επικοινωνία μεταξύ των μελών τους αλλά κυρίως τους προσδίδουν τη βασική ιδιότητα της σταθερότητας (robustness).

– Συντελεστής ομαδοποίησης

Ο συντελεστής ομαδοποίησης (clustering coefficient) ποσοτικοποιεί την τάση των κόμβων σε ένα δίκτυο να συγκεντρωθούν μαζί. Είναι μια πολύ σημαντική στατιστική ιδιότητα που σχετίζεται με το βαθμό στον οποίο οι κόμβοι του δικτύου τείνουν να σχηματίζουν τοπικά υποδίκτυα. Ο συντελεστής ομαδοποίησης ενός κόμβου καθορίζεται από την αναλογία των συνδέσεων μεταξύ των κόμβων εντός της άμεσης γειτονίας του, διαιρούμενος με τον αριθμό των συνδέσεων που θα μπορούσαν ενδεχομένως να υπάρχουν μεταξύ τους.

Στα πραγματική δίκτυα ο συντελεστής ομαδοποίησης είναι συνήθως πολύ μεγαλύτερος σε σχέση με τυχαία δίκτυα ίσου αριθμού κόμβων και ακμών [5]. Στην πράξη, μια υψηλή τιμή συντελεστή ομαδοποίησης, σε σύγκριση με τυχαίο δίκτυο του ίδιου μεγέθους, δείχνει ότι υπάρχουν ομάδες (συστάδες) κόμβων πυκνά συνδεδεμένες μεταξύ τους, αλλά με λίγες συνδέσεις με άλλες ομάδες.

– Φαινόμενο μικρού-κόσμου

Ένα δίκτυο μικρού-κόσμου (small world) [98, 150] είναι ένα δίκτυο όπου ο αριθμός των βημάτων που απαιτούνται για να επισκεφτεί κάποιος δύο τυχαία επιλεγμένους κόμβους αυξάνεται αρκετά αργά ως συνάρτηση του αριθμού των κόμβων του δικτύου. Με άλλα λόγια, παρά το συχνά μεγάλο μεγέθους τους, υπάρχει μια σχετικά σύντομη διαδρομή μεταξύ δύο οποιωνδήποτε κόμβων. Έτσι στα δίκτυα μικρού-κόσμου, ενώ οι περισσότεροι κόμβοι δεν γειτνιάζουν, είναι εφικτή η μετάβαση από οποιονδήποτε κόμβο σε οποιονδήποτε άλλο με μικρό αριθμό βημάτων.

Το φαινόμενο μικρού-κόσμου, γνωστό επίσης και ως 'έξι βαθμοί διαχωρισμού' (six degrees of separation), απαντάται σε πληθώρα δικτύων πραγματικού κόσμου π.χ., σιδηροδρομικά δίκτυα [131], μεταβολικά δίκτυα [148], τα δίκτυα των νευρώνων του εγκεφάλου [15], δίκτυα τροφικών πλεγμάτων [101] και το World Wide Web [80].

Τα δίκτυα μικρού-κόσμου είναι πιο ανθεκτικά σε διαταραχές από άλλες αρχιτεκτονικές δικτύων και η επικράτηση της αρχιτεκτονικής μικρού-κόσμου σε βιολογικά συστήματα εκλαμβάνεται ως ένα εξελικτικό πλεονέκτημα μιας τέτοιας αρχιτεκτονικής, καθώς τα δίκτυα μικρού-κόσμου μπορούν να θεωρηθούν ως συστήματα που είναι ιδιαιτέρως αποδοτικά, τόσο σε καθολικό, όσο και σε τοπικό επίπεδο, όσον αφορά την δυναμική των διεργασιών που συντελούνται σε αυτά π.χ., του πόσο αποτελεσματικά ανταλλάσσονται πληροφορίες μέσω του δικτύου [83].

– Αχίλλειος πτέρνα των πολύπλοκων δικτύων

Ένα ενδιαφέρον φαινόμενο των πολύπλοκων δικτύων είναι η 'αχίλλειος πτέρνα' τους: ανθεκτικότητα έναντι ευθραυστότητας [149]. Σε ένα δίκτυο ελεύθερης κλίμακας μικρού-κόσμου η διαγραφή ενός τυχαίου κόμβου σπάνια προκαλεί μια δραματική αύξηση στο βραχύτερο

μήκος διαδρομής. Αντίθετα, σε ένα τυχαίο δίκτυο, στο οποίο όλοι οι κόμβοι έχουν περίπου τον ίδιο αριθμό συνδέσεων, διαγράφοντας ένα οποιοδήποτε τυχαίο κόμβο θα αυξηθεί ελαφρώς σημαντικά το μήκος της μέσης συντομότερης διαδρομής. Με αυτή την έννοια, τα τυχαία δίκτυα είναι ευάλωτα σε τυχαίες διαταραχές, ενώ τα δίκτυα ελεύθερης κλίμακας μικρού-κόσμου είναι ανθεκτικά. Ωστόσο, τα δίκτυα ελεύθερης κλίμακας μικρού-κόσμου είναι ιδαιτέρως ευάλωτα σε στοχευμένη επίθεση των κόμβων, π.χ., με φθίνουσα σειρά βαθμών κόμβων, οδηγούμενα σε καταστροφική αποτυχία [6, 32].

2.1.2 Τεχνικές Διαφοροποίησης

Τεχνικές διαφοροποίησης έχουν προταθεί ως μέσο αντιμετώπισης της ασάφειας και του πλεονασμού, σε διάφορα προβλήματα και περιβάλλοντα, π.χ. διαφοροποίηση ιστορικών αρχείων [134], διαφοροποίηση αποτελεσμάτων συστάσεων [160], διαφοροποίηση αποτελεσμάτων αναζήτησης εικόνων [136], χρησιμοποιώντας μια πληθώρα αλγορίθμων και προσεγγίσεων, π.χ. αλγόριθμοι μάθησης [116], αλγόριθμοι προσέγγισης [92], διαφοροποιήσεις του pagerank[156] και πιθανότητες υπό όρους [28].

Οι χρήστες των μηχανών αναζήτησης χρησιμοποιούν συνήθως ερωτήματα με βάση λέξεις-κλειδιά για να εκφράσουν τις πληροφοριακές τους ανάγκες. Τα ερωτήματα αυτά είναι συχνά μη σαφώς προσδιορισμένα ή διφορούμενα σε κάποιο βαθμό [33]. Διαφορετικοί χρήστες που θέτουν ακριβώς το ίδιο ερώτημα ενδέχεται να έχουν πολύ διαφορετικές προθέσεις. Ταυτόχρονα, τα έγγραφα που ανακτώνται από ένα σύστημα ανάκτησης πληροφοριών μπορεί να αντανακλούν περιττές πληροφορίες. Η διαφοροποίηση αποτελεσμάτων αναζήτησης, η οποία στοχεύει στη διευκόλυνση των χρηστών κατά την αναζήτηση πληροφορίας, συνίσταται στην ανά-ταξινόμηση των αποτελεσμάτων και/ή στη συλλογή ενός περιορισμένου αριθμού αποτελεσμάτων, με τέτοιο τρόπο ώστε τα πρώτα αποτελέσματα που συλλέγονται να είναι όσο το δυνατόν πιο ετερογενή μεταξύ τους. Με τον τρόπο αυτό οι χρήστες θα ανακτήσουν αποτελέσματα τα οποία καλύπτουν διαφορετικές οπτικές γωνίες της πληροφοριακής ανάγκης τους. Γενικά, το πρόβλημα της επιλογής ποικιλόμορφων/ετερογενών αποτελεσμάτων ορίζεται ως: δοθέντος ενός συνόλου αποτελεσμάτων N, βρείτε ένα υποσύνολο $S \subset N$, ώστε με κάποιο κριτήριο ποικιλομορφίας να μεγιστοποιείται η ποικιλομορφία των αποτελεσμάτων στο S.

Η δημοσιευμένη βιβλιογραφία για τη διαφοροποίηση των αποτελεσμάτων αναζήτησης εξετάζεται στα [127, 37]. Οι ερευνητικές προσεγγίσεις που έχουν προταθεί ορίζουν γενικά τρεις βασικές κατηγορίες διαφοροποίησης αποτελέσματος βάσει συγκεκριμένων χαρακτηριστικών: (α) περιεχόμενο (content-based), τα διαφοροποιημένα αντικείμενα επιλέγονται βάσει τιμών περιεχομένου, (β) καινοτομία (nobelty-based), τα διαφοροποιημένα αντικείμενα περιέχουν νέα πληροφορία σε σχέση με εκείνα που προηγούνται, και (γ) κάλυψη (coverage-based), τα διαφοροποιημένα αντικείμενα καλύπτουν όσο δυνατόν περισσότερες θεματικές περιοχές. Οι προσεγγίσεις βασισμένες στο περιεχόμενο αντιμετωπίζουν το πρόβλημα της διαφοροποίησης ως μορφή του προβλήματος της p-διασποράς (p-dispersion) [40], όπου το ζητούμενο είναι η επιλογή, από ένα σύνολο n δεδομένων σημείων, p εξ΄ αυτών, ώστε η ελάχιστη, ανά ζεύγη, απόσταση των σημείων p να μεγιστοποιείται. Το πρόβλημα της p-διασποράς έχει μελετηθεί

στον κλάδο της επιχειρησιακής έρευνας για την χωροθέτηση εγκαταστάσεων που θα πρέπει να είναι όσο το δυνατό πιο απομακρυσμένες π.χ., πυρηνικά εργοστάσια.

Η πειραματική αξιολόγηση της αποτελεσματικότητας των μεθόδων ανάκτησης/ κατάταξης αποτελεί ανοικτή επιστημονική περιοχή. Χαλαρώνοντας την βασική υπόθεση της ανάκτησης πληροφοριών, ότι κάθε ερώτημα χρήστη αντιπροσωπεύει μια συγκεκριμένη ανάγκη πληροφόρησης, η αξιολόγηση καθίσταται περιπλοκότερη. Η αξιολόγηση των τεχνικών διαφοροποίησης βασίζεται στην θεώρηση των πληροφοριακών απαιτήσεων ως πολλαπλές πτυχές του ερωτήματος.

Μία από τις πλέον καθιερωμένες μεθοδολογίες αξιολόγησης της ανάκτησης, μακριά από τις ιδιαιτερότητες των μεμονωμένων χρηστών, βασίζεται στην αξιολόγηση της συνάφειας εμπειρογνωμόνων για την δημιουργία δεδομένων δοκιμών (evaluation benchmark) [126]. Τα δεδομένα δοκιμών αποτελούνται από τρία συστατικά: μια συλλογή εγγράφων, ένα σύνολο ερωτημάτων και ένα σύνολο εκτιμήσεων σχετικότητας, οι οποίες αντιστοιχίζουν έκαστο ερώτημα με τα συναφή με αυτό έγγραφα της συλλογής.

2.2 Επισκόπηση

Η ενότητα αυτή παρουσιάζει τις σημαντικότερες ερευνητικές προσεγγίσεις στις θεματικές περιοχές της παρούσης εργασίας. Αναλυτικότερα, στην υπό-ενότητα 2.2.1 παρουσιάζουμε μοντέλα και συστήματα που έχουν προταθεί για την αναπαράσταση και διαχείριση του δικαίου και στην υπό-ενότητα 2.2.2 παρουσιάζουμε δίκτυα νομικών παραπομπών. Τέλος στην υπό-ενότητα 2.2.3 εστιάζουμε σε και στην υπό-ενότητα 2.2.4 σε ανάλυση και διαφοροποίηση καταχωρήσεων χρηστών.

2.2.1 Αναπαράσταση Δικαίου

Στην υπό-ενότητα αυτή εστιάζουμε το ενδιαφέρον μας σε μοντέλα αναπαράστασης νομικής πληροφορίας, σε θέματα μοντελοποίησης νομικών πηγών, σε θέματα ανίχνευσης και επίλυση των νομικών παραπομπών και τέλος σε συστήματα διαχείρισης νομικών πηγών και νομικής γνώσης

Μοντέλα Αναπαράστασης Νομικής Πληροφορίας

Ένα μοντέλο αναπαράστασης νομικής γνώσης είναι απαραίτητο για την δόμηση των νομικών εγγράφων, την αναπαράσταση των μεταδεδομένων και συνδέσεων, και την μορφοποίηση νομικών εγγράφων. Αρκετά νομικά πρότυπα αναπτύχθηκαν τα τελευταία χρόνια. Μεταξύ των πιο ευρέως διαδεδομένων είναι τα MetaLex [22], NormeInRete [93], AkomaNtoso [14], η νομοθετική σήμανση των Ηνωμένων Πολιτειών USLM[1], η Νομοθεσία Ηνωμένου Βασιλείου UK[2] και Lexml Brasil [88]. Μια συγκριτική ανάλυση των κύριων προτύπων για τα νομοθετικά

[1] `https://github.com/usgpo/uslm`
[2] `http://www.legislation.gov.uk/developer/formats/xml`

έγγραφα, υπογραμμίζοντας συγκεκριμένα πλεονεκτήματα και αδυναμίες κάθε προτύπου, σε ευρωπαϊκό και παγκόσμιο επίπεδο παρουσιάζεται στο [90]. Σε αυτή την εργασία χρησιμοποιούμε το σχήμα Akoma Ntoso ως προς μοντέλο αναπαράστασης των νομικών εγγράφων, παρέχοντας επεκτάσεις για να αναπαραστήσουμε τις πηγές δικαίου στην Ελλάδα και τα μεταδεδομένα.

Μοντελοποίηση Νομικών Πηγών

Η έρευνα στο πεδίο της νομικής πληροφορικής επικεντρώνεται σε δύο προσεγγίσεις ως προς την κατασκευή μιας μηχανικά αναγνώσιμης μορφής αναπαράστασης των νομικών πηγών: δημιουργία μηχανικά αναγνώσιμης μορφής ταυτόχρονα με την δημιουργία του κειμένου και εκ των υστέρων. Η πρώτη προσέγγιση απαιτεί τη χρήση εξειδικευμένου λογισμικού επεξεργασίας, όπως παρουσιάζεται στα [109] και [1], όπου ο συντάκτης υποχρεούται να παρέχει πληροφορίες σχετικά με το κείμενο τη στιγμή της γραφής. Απαιτήσεις για ένα περιβάλλον σύνταξης νομικών πηγών με τεχνολογίες του σημασιολογικού ιστού παρουσιάζονται στο [143].

Σε αυτή την εργασία εστιάζουμε στην δεύτερη προσέγγιση της δημιουργίας ενός μοντέλου μετά τη δημιουργία του κειμένου, προσέγγιση που είναι επίσης κατάλληλη για υφιστάμενες νόμιμες πηγές. Χρησιμοποιεί αυτοματοποιημένες μεθόδους μετάφρασης νομικών πηγών, οι οποίες επικεντρώνονται σε επαναλαμβανόμενα μοτίβα που εμφανίζονται στο κείμενο. Στο πλαίσιο αυτό, η ανίχνευση δομής σημαίνει τον προσδιορισμό των διαφόρων τμημάτων του εγγράφου, όπως τα κεφάλαια, τα άρθρα και τις παραγράφους, και στη συνέχεια την επισήμανση με μεταδεδομένα αυτών. Ένας αναλυτής για την αυτόματη δόμηση ιταλικών νομοθετικών εγγράφων παρουσιάζεται στο [10]. Η προσέγγισή τους βασίζεται στην έκφραση των μορφών εισόδου σε hidden markov models. Αντίθετα στην εργασία μας χρησιμοποιείται γραμματική χωρίς πλαίσιο CFG[3] για την αποτύπωση της δομής και του περιεχομένου των νομικών πηγών. Οι συγγραφείς του [66] προτείνουν μια μέθοδο για την ανάλυση δομής εγγράφων με συντακτικό μοντέλο και αναλυτές για ιαπωνικές νομικές πηγές, είναι πιο κοντά στη δική μας. Σε αντίθεση με το [66], η μέθοδος μας δεν βασίζεται σε κανόνες PEG[4] [45], που δεν μπορούν να είναι διφορούμενοι, αλλά σε γραμματική χωρίς πλαίσιο, που επιτρέπει την ευελιξία όσον αφορά την ασάφεια σε όρους predicates [112].

Νομικές παραπομπές Έχουν προταθεί διάφορες μέθοδοι για την ανίχνευση και επίλυση των αναφορών μεταξύ των πηγών δικαίου [35, 105]. Η θεωρία των δικτύων εφαρμόστηκε επίσης στον τομέα του δικαίου για την κατασκευή δικτύων νομικών παραπομπών, αξιολογώντας τη συνάφεια των δικαστικών αποφάσεων [47] και βοηθώντας στη συνοπτική παρουσίαση δικαστικών αποφάσεων [51]. Στην παρούσα χρησιμοποιούμε αυτόματες μεθόδους για την ανίχνευση και επίλυση των αναφορών μεταξύ πηγών του νόμου και επίσης αξιοποιούμε το δίκτυο νομικών παραπομπών για να βοηθήσουμε την πλοήγηση των χρηστών και να διαφοροποιήσουμε τα αποτελέσματα αναζήτησης.

[3] `https://en.wikipedia.org/wiki/Context-free_grammar`
[4] `https://en.wikipedia.org/wiki/Parsing_expression_grammar`

Συστήματα Διαχείρισης Νομικών Πηγών και Νομικής Γνώσης

Η πρώιμη αισιοδοξία με την εμφάνιση των πρώτων έμπειρων νομικών συστημάτων έχει ξεθωριάσει για διάφορους λόγους [84]. Η έρευνα επικεντρώνεται σήμερα περισσότερο στον τομέα των συστημάτων διαχείρισης νομικής γνώσης. Το Eunomos [21] είναι ένα νομικό σύστημα διαχείρισης γνώσης που χρησιμοποιεί οντολογίες για την ταξινόμηση των νομικών πηγών. Σε σύγκριση με το προτεινόμενο σύστημα, δεν ακολουθεί το πρότυπο Linked Data Platform και δεν παρέχει προηγμένες υπηρεσίες αναζήτησης. Ο διακομιστής εγγράφων MetaLex [62] παρέχει νομικά έγγραφα ως συνδεδεμένα δεδομένα. Εφαρμόζει έναν γενικό μηχανισμό μετατροπής από μία δομημένη μορφή XML, σε μία άλλη CEN MetaLex. Αντίθετα, το προτεινόμενο σύστημα μοντελοποιεί αυτόματα τις νόμιμες πηγές σε XML και επιπρόσθετα προσδιορίζει και κάνει resolve τις νομικές παραπομπές. Σε μια παρόμοια προσέγγιση, το [50] αποσκοπεί στη δημοσίευση της φινλανδικής νομοθεσίας σε μορφή ανοιχτών συνδεδεμένων δεδομένων, μοντελοποιώντας νόμους και δικαστικές αποφάσεων σε μορφή RDF, εστιάζοντας περισσότερο στην επαναχρησιμοποίηση των δεδομένων σε διάφορες υπηρεσίες mash-up, παρά σε ένα σύστημα διαχείρισης νομικής γνώσης.

2.2.2 Δίκτυα Νομικών Παραπομπών

Η ανάλυση βιβλιογραφικών παραπομπών έχει χρησιμοποιηθεί στον τομέα της νομοθεσίας για την κατασκευή δικτύων παραπομπών σε δικαστικές υποθέσεις. Το Αμερικανικό νομικό σύστημα έχει μελετηθεί περισσότερο προς την κατεύθυνση αυτή. Στην ελληνική έννομη τάξη, όπως και στις υπόλοιπες έννομες τάξεις της Ευρώπης, που έλκουν το νομικό τους σύστημα από το Ρωμαϊκό Δίκαιο, οι πηγές των κανόνων δικαίου που διέπουν την έννομη τάξη είναι πολύ συγκεκριμένες. Αυτά τα συστήματα δικαίου αποκαλούνται civil law. Σε αντίθεση με το Αγγλοσαξονικό δίκαιο (common law) οι αποφάσεις των δικαστηρίων δεν αποτελούν πηγή του Δικαίου. Τα δικαστήρια ερμηνεύουν του ισχύοντες νόμους, οι δε αποφάσεις τους αποτελούν σπουδαία πηγή ερμηνείας του Δικαίου.

Οι συγγραφείς του [47] προτείνουν μεθόδους για τον προσδιορισμό των πιο κεντρικών αποφάσεων του Ανωτάτου Δικαστηρίου των ΗΠΑ. Οι ίδιοι συγγραφείς στο [48] μελετούν πως το δεδικασμένο είχε αλλάξει με την πάροδο του χρόνου στη νομολογία του Ανωτάτου Δικαστηρίου των ΗΠΑ, προκειμένου να προσδιορίσουν τα σημαντικότερα δεδικασμένα που περιέχονται στο Αμερικάνικο δίκαιο. Το δεδικασμένο αποτελεί νομικό κανόνα, που προέρχεται από το Αγγλικό δίκαιο, και ενθαρρύνει τους δικαστές να ακολουθήσουν προηγούμενη δικαστική απόφαση, η οποία αποτελεί 'νόμο' για την έκδοση απόφασης σε παρόμοια υπόθεση.

Σε αντίθεση το [144] καταλήγει στο συμπέρασμα ότι οι αλγόριθμοι κατάταξης [82] νομικών πηγών, που χρησιμοποιήθηκαν σε προηγούμενες έρευνες, όπως n-degree, HITS και PageRank, μπορεί να μην είναι οι πλέον κατάλληλοι για τη μέτρηση της νομικής βαρύτητας (legal authority). Για τον υπολογισμό της τελευταίας, εξωτερικοί παράγοντες, όπως ο αριθμός των δικαστών στην εκδίκαση της υπόθεσης ή το μέγεθος της απόφασης θα πρέπει να συνεκτιμηθούν. Ο ίδιος ερευνητής προτείνει ένα μοντέλο για την αυτοματοποιημένη αξιολόγηση της νομολογίας [145], το οποίο ενσωματώνει στοιχεία από τη δημοσίευση και τις

παραπομπές (citation) των δικαστικών υποθέσεων για την εκτίμηση της νομικής βαρύτητας των αποφάσεων.

Στο [135] διαπιστώνεται ότι το δίκτυο των αποφάσεων του Ανώτατου Δικαστηρίου των ΗΠΑ ακολουθεί κατανομή νόμου δύναμης (power-law). Στο [157] περιγράφεται ένα διαδραστικό σύστημα πλοήγησης που επιτρέπει την πλοήγηση σε σημασιολογικά νομικά δίκτυα παραπομπών και φανερώνει τις συσχετίσεις των παραπομπών.

Ωστόσο, οι μελέτες που προαναφέρθηκαν επικεντρώνονται σε συστήματα δικαίου βασιζόμενα στο common law: ένα νομικό σύστημα που αναπτύσσεται από τους δικαστές με αποφάσεις των δικαστηρίων και το οποίο διαφέρει από το civil law, που χρησιμοποιείται στην Ευρωπαϊκή Ένωση.

Στο [152] γίνεται ποσοτικοποίηση της πολυπλοκότητας της νομολογίας μέσω ανάλυσης του δικτύου παραπομπών, με ένα δείγμα 15.053 αποφάσεων του Ολλανδικού Ανώτατου Δικαστηρίου. Οι συγγραφείς επαληθεύουν τα αποτελέσματα του [47] για το δίκτυο παραπομπών του δειγματοληπτημένου νομικού συστήματος, δηλαδή ότι το δίκτυο παραπομπών του Ολλανδικού Ανώτατου Δικαστηρίου ακολουθεί κατανομή νόμου δύναμης (power-law). Με παρόμοιο τρόπο, η πολυπλοκότητα του Γαλλικού νομικού κώδικα αναλύεται στο [94], όπου, χρησιμοποιώντας ένα δείγμα από πενήντα δύο νομικούς κώδικες, εντοπίζονται διαρθρωτικές ιδιότητες του Γαλλικού νομικού κώδικα.

Το δεδικασμένο στα διεθνή δικαστήρια μελετάται στο [91]. Οι συγγραφείς χρησιμοποίησαν τεχνικές ανάλυσης δικτύου για να εξάγουν παραπομπές από το Ευρωπαϊκό Δικαστήριο για τα ανθρώπινα δικαιώματα και να καταλήξουν στο συμπέρασμα ότι τα διεθνή και τα εγχώρια δικαστήρια επανεξέτασης αναπτύσσουν τις αποφάσεις τους με παρόμοιο τρόπο. Κατά παρόμοιο τρόπο, το Διεθνές Ποινικό Δικαστήριο [137], το ιταλικό συνταγματικό δικαστήριο [2] εξετάστηκαν χρησιμοποιώντας τεχνικές ανάλυσης δικτύου. Τέλος, ένα σύνολο εργαλείων που επιτρέπει στους νομικούς μελετητές να εφαρμόσουν τεχνικές ανάλυσης δικτύου με σκοπό την οπτική παρουσίαση της νομολογίας της Ε.Ε. παρουσιάζεται στο [86].

Σε όλες τις παραπάνω μελέτες, αποδεικνύεται η αποτελεσματικότητα των τεχνικών ανάλυσης δικτύου στη νομολογία. Τα δίκτυα παραπομπών της νομολογίας περιέχουν πολύτιμες πληροφορίες, που μπορούν να μετρήσουν την βαρύτητα μιας δικαστικής απόφασης [47], να προσδιορίσουν τα σημαντικότερα δεδικασμένα [48] ή ακόμα και να προβλέψουν τις δικαστικές αποφάσεις που θα λάβουν τις περισσότερες αναφορές [145].

Ωστόσο, η ανάλυση του δικτύου παραπομπών στην νομολογία παρέχει πληροφορίες για μια μόνο διάσταση. Οι ακμές στο δίκτυο είναι του ίδιου τύπου, αναφορές μεταξύ των εγγράφων. Στο δίκαιο όμως υπάρχουν πολλά και ετερογενή δίκτυα, το καθένα από τα οποία αντιπροσωπεύει ένα συγκεκριμένο είδος συσχέτισης, και κάθε είδους συσχέτιση διαδραματίζει ξεχωριστό ρόλο στο νομικό δόγμα. Με τον τρόπο αυτό, προκειμένου να κατασκευάσουμε ένα μοντέλο δικτύου που να προσομοιώνει το δίκαιο με αρκετά αποτελεσματικό τρόπο, θα πρέπει να λάβουμε υπόψη την πολλαπλής κλίμακας δομή του δικαίου. Διακριτά χαρακτηριστικά, όπως η ιεράρχηση των πηγών του δικαίου, ή τα διάφορα είδη σχέσεων μεταξύ των νομικών πηγών θα πρέπει να ενσωματωθούν σε ένα μοντέλο, όπως αυτό που προτείνουμε στην παρούσα εργασία.

2.2.3 Τεχνικές Διαφοροποίησης Αποτελεσμάτων Αναζήτησης και Ανάκτησης Πληροφορίας σε Νομικές Πηγές

Στην ενότητα αυτή αρχικώς παρουσιάζουμε τεχνικές διαφοροποίησης αποτελεσμάτων αναζήτησης και στην συνέχεια, καθώς δεν υπάρχουν μέχρι και σήμερα εργασίες που να ασχολούνται με το πρόβλημα της διαφοροποιημένης ανάκτησης νομικής πληροφορίας, εστιάζουμε το ενδιαφέρον μας σε τεχνικές ανάκτησης νομικού περιεχομένου.

Οι τεχνικές διαφοροποίησης αποτελεσμάτων αναζήτησης έχουν προσελκύσει το ενδιαφέρον της ακαδημαϊκής κοινότητας στον τομέα ανάκτησης πληροφορίας τα τελευταία χρόνια. Στις εργασίες [37, 127] γίνεται εκτενής ανασκόπηση θεμελιωδών εργασιών. Μια από τις πρώτες εργασίες για την διαφοροποίηση, η μέγιστη οριακή σχετικότητα (MMR), παρουσιάζεται στο [25]. Η μέγιστη οριακή σχετικότητα έχει ως στόχο να μεγιστοποιήσει την συνάφεια ελαχιστοποιώντας παράλληλα την ομοιότητα του αποτελέσματος με τα προηγουμένως ανακτηθέντα αποτελέσματα. Με τον τρόπο αυτό τα αποτελέσματα της αναζήτησης επανα-κατατάσσονται συνδυάζοντας δύο επιμέρους συναρτήσεις: την ομοιότητά τους με το ερώτημα και την μεταξύ των αποτελεσμάτων ομοιότητά. Στο [58] εισάγεται ένα σύνολο από αξιώματα διαφοροποίησης και αποδεικνύεται ότι είναι αδύνατο κάποιος αλγόριθμος διαφοροποίησης να ικανοποιήσει το σύνολο των αξιωμάτων. Επιπλέον, δεδομένου ότι δεν υπάρχει ενιαία αντικειμενική συνάρτηση που να είναι κατάλληλη για κάθε τομέα εφαρμογής, οι συγγραφείς προτείνουν τρεις συναρτήσεις - στόχους διαφοροποίησης και αλγόριθμους διαφοροποίησης, τους οποίους και χρησιμοποιούμε στην εργασία μας.

Σε μια άλλη προσέγγιση, ερευνητές χρησιμοποίησαν εξωτερικές πηγές γνώσης για να διαφοροποιήσουν τα αποτελέσματα αναζήτησης. Στο [128] εισάγεται ένα πλαίσιο διαφοροποίησης, όπου οι διάφορες πτυχές ενός δεδομένου ερωτήματος αντιπροσωπεύονται από όρους δευτερογενών ερωτημάτων και τα έγγραφα ταξινομούνται με βάση τη συνάφεια τους με κάθε υπό-ερώτημα. Στο [3] προτείνεται ένας στόχο διαφοροποίησης που προσπαθεί να μεγιστοποιήσει την πιθανότητα εξεύρεσης σχετικού εγγράφου στις $top - k$ θέσεις, δοθείσας κατηγοριοποίησης των ερωτημάτων και των εγγράφων. Τέλος το [65] οργανώνει τις προθέσεις των χρηστών σε μια ιεραρχική δομή και προτείνει ένα πλαίσιο διαφοροποίησης για να αξιοποιήσει αυτή την ιεραρχική κατηγοριοποίηση.

Η βασική διαφορά μεταξύ αυτών των εργασιών και εκείνων που χρησιμοποιούνται στην εργασία μας είναι ότι δεν βασιζόμαστε σε εξωτερικές πηγές γνώσεων, π.χ. κατηγοριοποιήσεις/ταξινομίες, αρχεία καταγραφής ερωτημάτων για τη δημιουργία διαφοροποιημένων αποτελεσμάτων. Τα ερωτήματα των χρηστών σπάνια είναι γνωστά εκ των προτέρων, επομένως οι πιθανοτικές μέθοδοι για τον υπολογισμό των εξωτερικών πληροφοριών δεν είναι μόνο δαπανηρές για τον υπολογισμό τους, αλλά έχουν και εξειδικευμένο τομέα εφαρμογής. Αντίθετα, αξιολογούμε μεθόδους που βασίζονται μόνο στη ενδογενή γνώση του χρησιμοποιούμενου νομικού σώματος και σε τιμές που υπολογίζονται χρησιμοποιώντας συναρτήσεις ομοιότητας (συνάφειας) και ποικιλομορφίας στα δεδομένα.

Για την αξιολόγηση των αποτελεσμάτων της διαφοροποίησης, στο [30] εισάγεται ένα πλαίσιο αξιολόγησης της καινούργιας πληροφορίας και της διαφοροποίησης. Στο [147] παρου-

σιάζεται μία μέθοδος για διαφοροποίηση δομημένων δεδομένων, όπου τα στοιχεία προς διαφοροποίηση δεν είναι έγγραφα, αλλά αντικείμενα με διακριτές ιδιότητες, δηλαδή εγγραφές σε ένα πίνακα μίας βάσης. Το [28] θεωρεί μία μετρική αξιολόγησης η οποία δίνει ποινή σε ένα μοντέλο ανάκτησης αποτελεσμάτων, μόνο αν δεν επιστρέφει καθόλου σχετικά αποτελέσματα. Με βάση αυτό, οι συγγραφείς προτείνουν μία μέθοδο στην οποία κάθε αποτέλεσμα επιλέγεται με βάση την πιθανότητα να (μην) είναι όμοιο με τα προηγουμένως επιλεγμένα αποτελέσματα.

Διάφορες διαστάσεις της συνάφειας στην ανάκτηση νομικών πληροφοριών, βάσει συγκεκριμένων χαρακτηριστικών των νομικών πηγών και του δικαίου εξετάζονται στο [146]. Η ανάκτηση νομικών πηγών, νομικού κειμένου, παραδοσιακά βασίζεται σε εξωτερικές πηγές γνώσης, όπως π.χ. οι θησαυροί και τα συστήματα ταξινόμησης και διάφορες τεχνικές παρουσιάζονται στο [99]. Επιτηρούμενες μέθοδοι μάθησης έχουν προταθεί για την ταξινόμηση των πηγών δικαίου σύμφωνα με τις νομικές έννοιες [18, 97, 59]. Οντολογίες και θησαυροί έχουν χρησιμοποιηθεί για τη διευκόλυνση της ανάκτησης πληροφοριών [129, 130, 53, 70] ή για την ανταλλαγή πληροφοριών μεταξύ υφισταμένων συστημάτων νομικής γνώσης [63]. Βασική διαφορά μεταξύ των προηγούμενων μεθόδων και εκείνων που χρησιμοποιούνται σε αυτή την εργασία είναι ότι δεν βασιζόμαστε σε εξωτερικές πηγές γνώσεων, π.χ. θησαυροί, αρχεία καταγραφής ερωτημάτων για τη δημιουργία διαφοροποιημένων αποτελεσμάτων. Τα ερωτήματα των χρηστών σπάνια είναι γνωστά εκ των προτέρων, επομένως οι πιθανοτικές μέθοδοι για τον υπολογισμό των εξωτερικών πληροφοριών δεν είναι μόνο δαπανηρές για τον υπολογισμό τους, αλλά έχουν και εξειδικευμένο τομέα εφαρμογής. Αντίθετα, χρησιμοποιούμε μεθόδους που βασίζονται ενδογενή γνώση (implicit knowledge) του σώματος των νομικών πηγών και υπολογιζόμενες τιμές, χρησιμοποιώντας συναρτήσεις ομοιότητας (συνάφειας) και ποικιλομορφίας στα δεδομένα.

Τεχνικές αυτόματης σύνοψης και περίληψης νομικών εγγράφων έχουν προταθεί για να καταστήσουν το κείμενο νομικών εγγράφων, κυρίως δικαστικών αποφάσεων, ευκολότερα προσβάσιμο [42, 43, 100]. Στην συγκεκριμένη εργασία χρησιμοποιούμε επίσης αλγόριθμους περιλήψεων, προσαρμόζοντας τους στο συγκεκριμένο πρόβλημα: στοχεύουμε στη μεγιστοποίηση της ποικιλομορφίας του αποτελέσματος αναζήτησης που εξυπηρετεί ένα δεδομένο ερώτημα. Οι αλγόριθμοι περιλήψεων κειμένων [LexRank [39] και Biased LexRank [106]] προτάθηκαν αρχικά για τον υπολογισμό της σχετικής σημασίας των κειμενικών μονάδων μέσα σε ένα έγγραφο για την υποβοήθηση εργασιών συνοπτικής τεκμηρίωσης. Θεωρούν ένα έγγραφο ως γράφο με κόμβους τις προτάσεις του εγγράφου και ακμές που σχηματίζονται μεταξύ των κόμβων με βάση την κειμενική ομοιότητα των προτάσεων. Για να χρησιμοποιήσουμε μια τέτοια προσέγγιση στο σενάριο διαφοροποίησης μας, θα πρέπει:

- να εισάγουμε ως προαπαιτούμενο το ερώτημα του χρήστη, καθώς στην προσέγγισή μας, χρησιμοποιούμε τα έγγραφα που βρίσκονται στην λίστα των N ανακτηθέντων εγγράφων, με βάση την συνάφεια, για ένα δεδομένο ερώτημα και

- να δημιουργήσουμε ένα γράφο χρησιμοποιώντας την αρχική λίστα των N ανακτηθέντων εγγράφων, θεωρώντας τα έγγραφα ως κόμβους, και ακμές που σχηματίζονται μεταξύ των κόμβων με βάση το βαθμό της ανά μεταξύ τους ομοιότητας.

Δίκτυα νομικών παραπομπών, όπως παρουσιάστηκε στην ενότητα 2.2.2, έχουν επίσης προταθεί στην βιβλιογραφία σε θέματα ανάκτησης νομικής πληροφορίας, όπως για παράδειγμα για τον προσδιορισμό των πιο κεντρικών αποφάσεων του Ανωτάτου Δικαστηρίου των ΗΠΑ [47] και των σημαντικότερων δεδικασμένων [48]. Στην συγκεκριμένη εργασία χρησιμοποιούμε επίσης γνώση που απορρέει από το δίκτυο νομικών παραπομπών, προσαρμόζοντας αλγόριθμους διαφοροποιημένης κατάταξης σε γράφους [DivRank [95] και Grasshopper [159]], με στόχο την κάλυψη περισσότερων εκφάνσεων δοθέντος ερωτήματος χρήστη. Για να ενσωματώσουμε τις προαναφερθείσες τεχνικές στο σενάριο διαφοροποίησης μας, θα πρέπει:

- να εισάγουμε ως προαπαιτούμενο το ερώτημα του χρήστη, καθώς οι εν λόγω μέθοδοι έχουν προταθεί ανεξάρτητα ερωτήματος χρήστη και
- να δημιουργήσουμε ένα γράφο χρησιμοποιώντας την αρχική λίστα των N ανακτηθέντων εγγράφων, με βάση τις νομικές παραπομπές μεταξύ των εγγράφων αυτών.

Τέλος, μια παρόμοια προσέγγιση με την εργασία μας περιγράφεται στο [4], όπου οι συγγραφείς χρησιμοποιούν τεχνικές ανάκτησης πληροφορίας για να προσδιορίσουν ποια τμήματα ενός νομικού εγγράφου τείνουν να αποκλίνουν (outliers). Η εργασία μας διαφέρει καθώς μεγιστοποιούμε τη ποικιλομορφία του συνόλου των αποτελεσμάτων, αντί να εντοπίζουμε αποκλίνοντα τμήματα εντός συγκεκριμένης νομικής πηγής.

2.2.4 Ανάλυση και Διαφοροποίηση Καταχωρήσεων Χρηστών

Αρκετές μελέτες σχετικά με την αναζήτηση καταχωρήσεων χρησιμοποιούν διαφορετικά χαρακτηριστικά και αλγόριθμους για την κατάταξη των καταχωρήσεων. Οι συγγραφείς του [67] ενσωματώνουν διάφορες παραμέτρους σε ένα Bayesian μοντέλο. Στο [158] χρησιμοποιούνται αλγόριθμοι learning to rank και ορίζεται η σχετικότητα ενός tweet ως η πιθανότητα re-tweet αυτού.

Οι περισσότερες από τις μελέτες σχετικά με την αναζήτηση καταχωρήσεων χρηστών επικεντρώνουν το ενδιαφέρον στην συνάφεια με το ερώτημα του χρήστη και παραμελούν την διαφοροποίηση των αποτελεσμάτων. Η διαφοροποίηση των αποτελεσμάτων αναζήτησης αντιμετωπίζεται στο [122] με βάση την μέγιστη οριακή σχετικότητα και συσταδοποίηση των tweets. Σε αντιδιαστολή, στην παρούσα εργασία ορίζουμε εξειδικευμένα κριτήρια διαφοροποίησης και εφαρμόζουμε ευριστικούς αλγόριθμους σχετικούς με το πρόβλημα p-διασποράς (p-dispersion)[40]. Μια μελέτη για την περίληψη γεγονότων [121], προτείνει μεθόδους για την εύρεση του πιο αντιπροσωπευτικού σύνολου των tweets για ένα γεγονός από μια αρχική σειρά tweets. Παρόλο που η περιληπτική παρουσίαση γεγονότων και η διαφοροποίηση των αποτελεσμάτων αναζήτησης μοιράζονται ένα κοινό απώτερο στόχο, την εξεύρεση ενός αντιπροσωπευτικού σύνολου tweets σχετικά με ένα θέμα, θα πρέπει να σημειώσουμε ότι το ερώτημα είναι γενικότερη έννοια από το γεγονός. Τέλος, το πρόβλημα της διαφοροποίησης των καταχωρήσεων χρηστών αποτελεί επίσης αντικείμενο του [29]. Οι συγγραφείς υπολογίζουν το μικρότερο υποσύνολο των καταχωρήσεων που καλύπτουν όλες τις άλλες σε σχέση με την διάσταση της κειμενικής ομοιότητας. Σε σύγκριση με το [29], η εργασία μας εστιάζεται στον εντοπισμό των καταχωρήσεων που αναφέρονται σε πολλαπλές διαστάσεις, ενώ η

[29] παρέχει λύση για μία μόνο διάσταση. Επιπλέον, αντί να εστιάζουμε στην κάλυψη του συνόλου των αποτελεσμάτων, στοχεύουμε στη μεγιστοποίηση της πολυμορφίας του συνόλου των αποτελεσμάτων.

Μια παρόμοια προσέγγιση περιγράφεται επίσης στο [154], όπου οι συγγραφείς παρουσιάζουν ένα σύστημα για διαδικτυακές ομάδες συζήτησης το οποίο προτείνει διάφορες απόψεις και επιτρέπει στον χρήστη να ρυθμίσει το βαθμό της ομοιότητας/ποικιλομορφίας των συστάσεων σε σχέση με τις απόψεις του. Σε αντιδιαστολή, η εργασία μας, χρησιμοποιεί διαφορετικά κριτήρια διαφοροποίησης, δεν απαιτεί ρητή ανάδραση από τους χρήστες και επιπρόσθετα διαφοροποιεί το περιεχόμενο με τρόπο καθολικό και όχι με βάση τις προσωπικές απόψεις του χρήστη.

Τεχνικές κατηγοριοποίησης προτείνονται στο [140] για την εκτίμηση της πιθανότητας σχολιασμού ενός άρθρου, ενώ στο [141] μοντελοποιούνται και συγκρίνονται κατανομές σχολίων άρθρων με βάση διάφορες πηγές ειδησεογραφίας. Με αυτόν τον τρόπο γίνεται πρόβλεψη για τον συνολικό αριθμό σχολίων ενός άρθρου, μετά από μια αρχική περίοδο σχολιασμού του. Στο [132] προτείνεται μια μεθοδολογία συστάσεων άρθρων σε χρήστες, που είναι πιθανό να σχολιαστούν από αυτούς. Η προτεινόμενη μέθοδος χρησιμοποιεί το περιεχόμενο των άρθρων και μοτίβα σχολιασμού άρθρων από χρήστες. Επίσης τα σχόλια των χρηστών χρησιμοποιούνται στο [87] για την δημιουργία ενός συστήματος συστάσεων άρθρων σε κοινωνικά δίκτυα. Με βάση το περιεχόμενο του άρθρου και τα σχόλια του, κατασκευάζονται θεματικά προφίλ με βάση τα οποία ανακτώνται σχετικά προς σύσταση άρθρα. Στο [61] μελετώνται τα σχόλια, οι υπερσύνδεσμοι και οι διασυνδέσεις μεταξύ blog, ενώ στο [36] μελετώνται οι ανάγκες σχολιασμού άρθρων από τους χρήστες και αναλύονται ποιοτικά τα σχόλια που αναρτώνται στην ιστοσελίδα μίας διαδικτυακής εφημερίδας. Στο [55] συσχετίζονται η δημοτικότητα ενός blog με το μοτίβο των σχολίων του, αποτιμάται η συνεισφορά των σχολίων στην προσβασιμότητα του blog και εκτιμάται το σύνολο των σχολίων blog στον παγκόσμιο ιστό. Οι συγγραφείς του [114] καταλήγουν ότι είναι δυνατή η εύρεση συγκεκριμένου πλήθους σχολίων που αντιπροσωπεύουν το αντίστοιχο άρθρο. Στο [64] προτείνεται μέθοδος για την παραγωγή περίληψης του άρθρου, χρησιμοποιώντας τα σχόλια αυτού.

Το συναίσθημα των σχολίων των χρηστών σε πολιτικά άρθρα αναγνωρίζεται στο [110] και εν συνεχεία χρησιμοποιείται για να προβλέψει τον πολιτικό προσανατολισμό τους, ενώ στο [79] γίνεται ανάλυση συναισθημάτων σε αναρτήσεις χρηστών και αναλύεται η επιρροή διαφόρων παραγόντων, όπως δημογραφικά στοιχεία, θεματικές κατηγορίες και χρονική στιγμή της ανάρτησης. Τέλος, η ικανοποίηση των χρηστών για πολιτικές γνώμες αξιολογείται στο [102], όπου παράλληλα υιοθετείται μια κατηγοριοποίηση των χρηστών σε εκείνους που αναζητούν ταυτόσημες με τις δικές τους απόψεις και εκείνους που αναζητούν ετερογενείς απόψεις.

Κεφάλαιο 3

Μοντελοποίηση και Διαχείριση Νομικών Πηγών

Στο κεφάλαιο αυτό μελετάμε θέματα μοντελοποίησης και διαχείρισης νομικών πηγών χρησιμοποιώντας τεχνολογίες του σημασιολογικού ιστού. Αρχικά προτείνουμε μια πρότυπη μέθοδο για την αυτόματη μοντελοποίηση και σημασιολογική αναπαράσταση νομικών πηγών [78], μέσω της δημιουργίας μιας γλώσσας συγκεκριμένου τομέα για τις νομικές πηγές. Στην συνέχεια, προτείνουμε την αρχιτεκτονική μιας πλατφόρμας διαχείρισης νομικών πηγών, η οποία έχει στόχο την βελτίωση της πρόσβασης σε νομικές πηγές παρέχοντας προηγμένες υπηρεσίες μοντελοποίησης, διαχείρισης και ανάκτησης νομικής πληροφορίας [77].

3.1 Κίνητρο και Συνεισφορά

Στην σημερινή εποχή, ως συνέπεια πρωτοβουλιών για ανοιχτά δεδομένα και ιδιαίτερα για ανοιχτά κυβερνητικά δεδομένα, παρατηρείται μια τεράστια αύξηση στον αριθμό των ιστοχώρων (portals) που παρέχουν νομικά έγγραφα στους ενδιαφερόμενους. Νομικές πηγές, στις οποίες είχε προηγούμενα πρόσβαση μόνο ένα εξειδικευμένο κοινό, είναι πλέον ελεύθερα διαθέσιμες στο διαδίκτυο. Παρόλο που αυτή η ανοικτή πρόσβαση υποστηρίζεται συνήθως από υπηρεσίες θεματικής πλοήγησης και αναζήτησης με λέξεις κλειδιά, οι νομικοί πόροι διατίθενται, κατά κύριο λόγο, σε μια σημασιολογικά φτωχή κειμενική αναπαράσταση, η οποία δεν αποτυπώνει τη δομή και τη νομική σημασιολογία των δεδομένων.

Οι πρόσφατες τεχνολογικές εξελίξεις στο σημασιολογικό ιστό [17], ένα σύνολο διασυνδεδεμένων δεδομένων, προσβάσιμων και επεξεργάσιμων από εφαρμογές, παρέχουν το υπόβαθρο για τη θεμελίωση του νομικού σημασιολογικού ιστού [16], όπου οι νομικές πηγές είναι κατανοητές από υπολογιστικά συστήματα, αναγνωρίσιμες μεταξύ των ιστοτόπων, διασυνδεδεμένες και επεξεργάσιμες σύμφωνα με τη νομική τους σημασία. Απαραίτητη προϋπόθεση για την εκπλήρωση των προαναφερθέντων αποτελεί η δόμηση των νομικών πηγών σε μια τυποποιημένη μορφή, η ύπαρξη ενός κοινού λεξιλογίου αναφοράς μεταξύ των εμπλεκομένων συστημάτων. Διάφορες πρωτοβουλίες, σε εθνικό και διεθνές επίπεδο, έχουν προτείνει προτύπα για την συντακτική και σημασιολογική αναπαράσταση των νομικών πηγών και των μεταδεδομένων τους

[22, 93, 14].

Η έκθεση με τίτλο 'παγκόσμιο ηλεκτρονικό κοινοβουλίο' για το έτος 2016 [120], διαπιστώνει ότι μόνο 49% των κοινοβουλίων διαθέτουν συστήματα διαχείρισης εγγράφων και μόλις 26% χρησιμοποιούν κάποιο πρότυπο του σημασιολογικού ιστού π.χ., μορφή XML ως μορφή διανομής νομικών εγγράφων. Οι νομικές πηγές προσφέρονται στον τελικό χρήστη, κατά βάση, σε μια φιλική και ανθρώπινα αναγνώσιμη μορφή, προσανατολισμένη στην παρουσίαση. Η χρήση αυτού του ιδιόκτητου και μη δομημένου μορφοτύπου, κυρίως PDF, καθιστά αδύνατη την καθιέρωση διαλειτουργικότητας μεταξύ των διαφόρων παρόχων νομικής πληροφορίας, μη επιτρέποντας την επαναχρησιμοποίηση του περιεχομένου και την διασύνδεση με αποθετήρια στον σημασιολογικό ιστό. Ταυτόχρονα, το πλήθος των διαθέσιμων νομικών δεδομένων δυσχεραίνει τόσο τους επαγγελματίες του νομικού χώρου όσο και τους πολίτες να εντοπίσουν σχετικούς και χρήσιμους νομικούς πόρους.

Με βάση την προαναφερθείσα έλλειψη διαθεσιμότητας νομικών εγγράφων σε δομημένο και τυποποιημένο μορφότυπο, σε αυτό το κεφάλαιο, παρουσιάζουμε μια πρότυπη μεθοδολογία για την αυτόματη μοντελοποίηση και σημασιολογική αναπαράσταση νομικών πηγών, από αδόμητες μορφές (Ενότητα 3.2). Η μεθοδολογία μας εκμεταλλεύεται τα κοινά χαρακτηριστικά των νομικών πηγών για τον εντοπισμό τμημάτων νομικών πληροφοριών, τα οποία και μοντελοποιεί και διασυνδέει σημασιολογικά. Εκφράσαμε την δομή των νομικών πηγών με τη μορφή ενός συνόλου συντακτικών κανόνων, δηλαδή μιας γλώσσας συγκεκριμένου τομέα για τις νομικές πηγές, η οποία και χρησιμοποιείται για τη δημιουργία ενός συντακτικού αναλυτή νομικών πηγών. Ο συντακτικός αναλυτής, επεξεργάζεται δοθέντα νομικά κείμενα, σύμφωνα με τους κανόνες σύνταξης που περιγράφονται από την ειδική γλώσσα, αναγνωρίζοντας τη δομή και τα μεταδεδομένα τους, μοντελοποιόντας τα στο σχήμα που χρησιμοποιείται για τη σημασιολογική αναπαράσταση των νομικών πόρων. Για το σκοπό αυτό χρησιμοποιούμε το Akoma Ntoso, ένα δημοφιλές νομικό μετα-σχήμα, το οποίο και προσαρμόζουμε για να μοντελοποιήσουμε τις Ελληνικές νομικές πηγές. Η αξιολόγηση της προσέγγισής μας , φανερώνει ότι ο μηχανισμός που υλοποιήσαμε εξάγει τη δομή και τα μεταδεδομένα του νομικών πηγών συνδυάζοντας υψηλή ακρίβεια με γραμμική απόδοση.

Η μέθοδος αυτόματης μοντελοποίησης αποτελεί μέρος ευρύτερης αρχιτεκτονικής που προτείνουμε σε αυτό το κεφάλαιο, για την διαχείριση νομικών πηγών (Ενότητα 3.3), υπό την ονομασία Solon. Ο κύριος στόχος του Solon είναι να βοηθήσει τους χρήστες να εντοπίσουν και να ανακτήσουν νομικά και κανονιστικά έγγραφα μέσα στο ακριβές πλαίσιο εννοιολογικής αναφοράς. Αποτελείται από πολλά διαφορετικά συστατικά (components), τα οποία επικοινωνούν και προσφέρουν υπηρεσίες με βάση την αρχιτεκτονική REST. Πρόκειται για μια εξελιγμένη πλατφόρμα διαχείρισης νομικών πηγών που λειτουργεί σε νομικά έγγραφα, τα οποία ανακαλύπτονται και συλλέγονται αυτόματα από δικτυακές πύλες μέσω εξειδικευμένων εργαλείων συγκομιδής. Δεδομένου ότι τα νομικά έγγραφα διαχέονται σε μη μηχαναγνώσιμες μορφές, είναι απαραίτητος ο αυτόματος μετασχηματισμός τους σε μορφή κατάλληλη για τη μοντελοποίηση νομικών πηγών, με σκοπό την δομική και σημασιολογική αναπαράσταση τους, τη διασύνδεσή τους βάση νομικών παραπομπών και την ταξινόμησή τους με ένα σύνολο κανόνων. Ο Solon αξιοποιεί την σημασιολογική αναπαράσταση των νομικών πηγών, προσφέροντας, μεταξύ άλ-

λων, εξελιγμένα αποτελέσματα αναζήτησης και επιτρέποντας στους χρήστες να οργανώσουν τις νομικές πληροφορίες σύμφωνα με τις ατομικές τους προτιμήσεις. Η αρχιτεκτονική του Solon αξιολογήθηκε σε περιβάλλον παραγωγής, δημόσιου φορέα, παρέχοντας στο ευρύ κοινό σημασιολογική πρόσβαση στην ελληνική φορολογική νομοθεσία.

3.2 Αυτόματη Μοντελοποίηση και Σημασιολογική Ανάλυση Νομικών Κειμένων

Σε αυτή την ενότητα εξετάζουμε τις απαιτήσεις σχετικά με τη δομή των ελληνικών νομικών πηγών και την μεθοδολογία μετασχηματισμού σε κανόνες συντακτικού και δημιουργίας του συντακτικού αναλυτή [1]. Επιπρόσθετα, παρουσιάζουμε τα αποτελέσματα της πειραματικής αξιολόγησης της μεθόδου μας.

3.2.1 Δομή Νομικών Πηγών

Οι νομικές πηγές βασίζονται σε σαφώς καθορισμένη λεκτική δομή διατύπωσης/διάρθρωσης και 'αυστηρή' γλώσσα σύνταξης. Κατευθυντήριες γραμμές και πρότυπα καλής νομοθετικής σύνταξης, τόσο σε εθνικό όσο και σε επίπεδο Ε.Ε. π.χ., Κοινός Πρακτικός Οδηγός για τα πρόσωπα που εμπλέκονται στην εκπόνηση της νομοθεσίας της Ευρωπαϊκής Ένωσης [2], έχουν θεσπίσει κοινούς μορφότυπους, βάση των οποίων και συντάσσονται τα περισσότερα νομικά έγγραφα. Σε αυτήν την εργασία ο όρος 'μορφότυπος' χρησιμοποιείται για να υποδηλώσει θέματα όπως η δομή, η αρίθμηση, η προτιμώμενη χρήση λέξεων, η γραμματική σύνταξη και πρότυπα προτάσεων.

Σε απλουστευμένη μορφή, μία νομική πηγή έχει την ακόλουθη δομή:

- Προοίμιο - *Εισαγωγικό μέρος*, που περιέχει πληροφορίες που μας επιτρέπουν να προσδιορίσουμε τον τύπο του εγγράφου καθώς και μεταδεδομένα του όπως τον τίτλο, τον αριθμό, μεταδεδομένα έκδοσης π.χ., η αρχή έκδοσης κ.λπ.

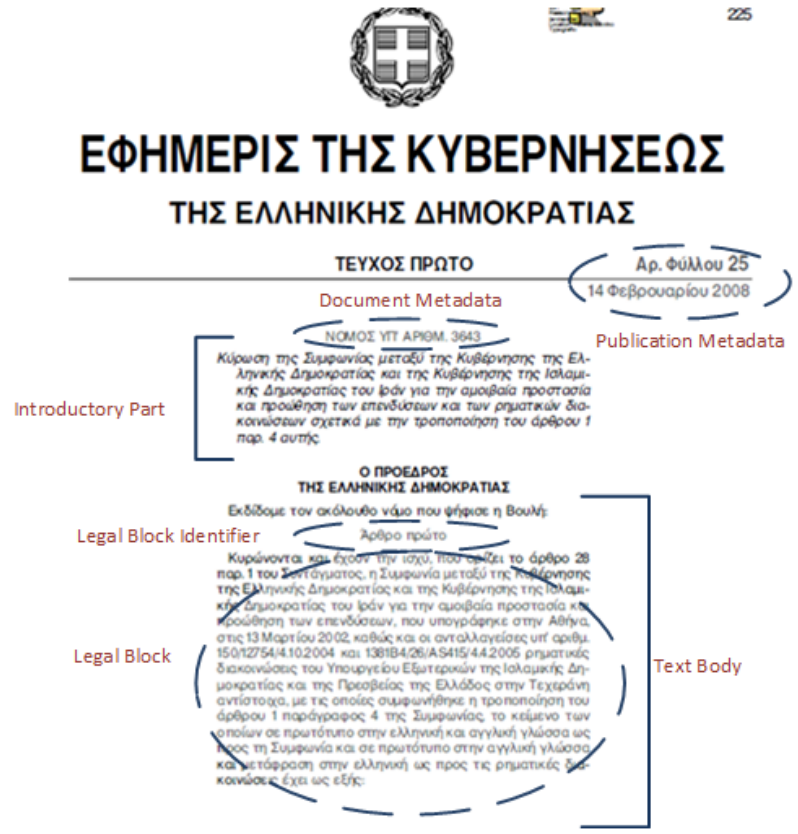
- Κύριο μέρος - *Σώμα κειμένου*, το οποίο αποτελεί το κύριο μέρος του εγγράφου. Το σώμα του κειμένου ακολουθεί διαφορετική διάρθρωση ανάλογα με τον τύπο του νομικού εγγράφου. Ωστόσο, ακολουθεί συνήθως μια καλά καθορισμένη ιεραρχική διάταξη μπλοκ κειμένου (*νομικά μπλοκ*), στα οποία τμήματα υψηλού επιπέδου ενός εγγράφου, όπως το κεφάλαιο ενός νόμου, περιέχουν άλλα μπλοκ, όπως άρθρα και παραγράφους.

- Τελικές διατάξεις, που περιέχει τύπους κλεισίματος, όπως ημερομηνία, υπογράφοντες κ.λπ.

Το Σχήμα 3.1 παρέχει ένα οπτικό βοήθημα της προαναφερθείσας δομής για ένα νόμο, όπου σημειώσαμε τα δομικά μέρη και τιμές μεταδεδομένων που απαντώνται.

[1] Συμπληρωματικό τεχνικό υλικό παρέχεται στο `https://github.com/mkoniari/LegalParse`

[2] `http://eur-lex.europa.eu/content/pdf/techleg/joint-practical-guide-2013-en.pdf`



Σχήμα 3.1: Επισκόπηση της δομής ενός νομικού εγγράφου με επισημειώσεις δομικών τμημάτων και μεταδεδομένων

3.2.2 Γλώσσα Ειδικού Σκοπού Περιγραφής Νομικών Πηγών

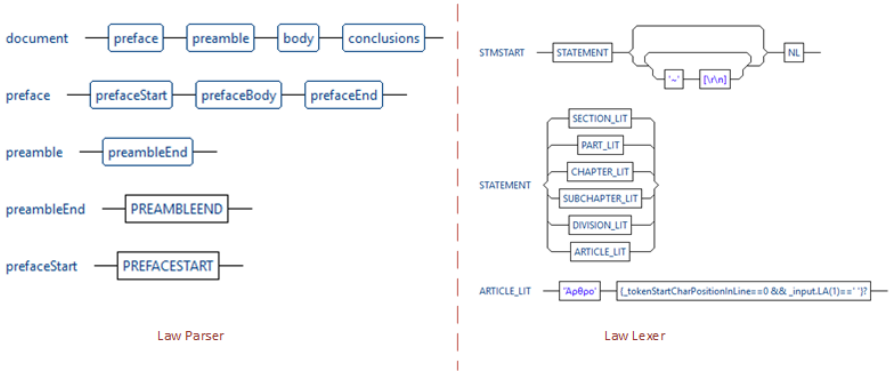
Το ειδικό πλαίσιο μοντελοποίησης (Domain-specific modeling) [41], αποτελεί μεθοδολογία της τεχνολογίας λογισμικού για την σχεδίαση και την ανάπτυξη συστημάτων, με χρήση γλωσσών ειδικού σκοπού, προσφέροντας ειδικά προσαρμοσμένες λύσεις για την αντιμετώπιση προβλημάτων σε συγκεκριμένο τομέα. Η δομή των νομικών πηγών, όπως περιγράφηκε προηγουμένως, μπορεί να εκφραστεί με βάση το ειδικό πλαίσιο μοντελοποίησης και να εξαχθούν κανόνες, με χρήση του μοντέλου γλώσσας ειδικού σκοπού (DSL) [111, 49]. Με τον τρόπο αυτό, κάνοντας χρήση ενός αφηρημένου μοντέλου νομικών πηγών, που περιγράφει τη συντακτική δομή τους, μπορούμε να κατασκευάσουμε ένα συντακτικό αναλυτή υψηλής γενικότητας και επεκτασιμότητας.

Μια γραμματική χωρίς συμφραζόμενα (context-free grammar - GFG) είναι μια τετράδα $G = (V, T, P, S)$ που αποτελείται από:

- V, το αλφάβητο από μη τερματικά σύμβολα (μεταβλητές)
- T το αλφάβητο τερματικών συμβόλων με $V \cap T \neq \varnothing$
- P, είναι ένα σύνολο κανόνων παραγωγής που αντιπροσωπεύουν τον αναδρομικό ορισμό της γλώσσας, δηλαδή ζεύγη $A \to \alpha$ με $\alpha \in (V \cup T)^*$ και $A \in V$

– S είναι το σύμβολο εκκίνησης που αντιπροσωπεύει τη γλώσσα που ορίζεται.

Η Εκτεταμένη Μορφή Μπάκους-Νάουρ (Extended Backus–Naur Form, EBNF) αποτελεί μια αρκετά συνηθισμένη τεχνική συμβολισμού για γραμματικές GFG. Κανόνες σύνταξης, σε μορφή EBNF, για την δομή του νομικού εγγράφου που παρουσιάζεται στο Σχήμα 3.1 παρουσιάζονται στο Σχήμα 3.2, όπου τα μη τερματικά σύμβολα body και conclusions έχουν οριστεί ξεχωριστά. Preface και preamble συνθέτουν το εισαγωγικό μέρος του εγγράφου και το body το κυρίως κείμενο. Το Σχήμα 3.2 παρέχει επίσης τμήμα των κανόνων γραμματικής για συντακτική (αριστερό μέρος) και λεκτική ανάλυση (δεξί μέρος). Ο συνδυασμός διαφορετικών κανόνων γραμματικής επιτρέπει τον προσδιορισμό του τύπου νομικού εγγράφου και των νομικών μπλοκ κειμένου που αυτό περιέχει.



Σχήμα 3.2: Επισκόπηση κανόνων γραμματικής νομικών πηγών

Εμπειρογνώμονες του νομικού τομέα μας βοήθησαν στον ορισμό των νομικών μπλοκ και των στοιχείων που αυτά μπορεί να περιέχουν καθώς και στις σχέσεις τους (π.χ. εμφύτευση, κληρονομική διαδοχή κ.λπ.) μέσα στα νομικά έγγραφα. Με βάση την καθορισμένη γραμματική και το σύνολο των κανόνων συντακτικού των νομικών πηγών μπορούμε με κάποια γεννήτρια κατασκευής συντακτικών αναλυτών να παράγουμε με αυτόματο τρόπο πηγαίο κώδικα για την συντακτική ανάλυση των νομικών εγγράφων. Γεννήτριες κατασκευής συντακτικών αναλυτών χρησιμοποιούνται κυρίως από μεταγλωττιστές (compilers) και διερμηνείς (interpreters) για να διαβάζουν τα αρχεία πηγαίου κώδικα που πρέπει να μεταγλωττιστεί ή να εκτελεστεί. Ωστόσο, μπορούν να χρησιμοποιηθούν και σε άλλες εφαρμογές, όταν υπάρχει η ανάγκη επεξεργασίας ή ερμηνείας δεδομένων που ακολουθούν τον φορμαλισμό κάποιας τυπικής γλώσσας, όπως στην περίπτωσή μας.

Μεταξύ των δημοφιλέστερων γεννητριών συγκαταλέγεται το Another Tool for Language Recognition, ANTLR[3]. Το ANTLR δέχεται ως είσοδο οποιαδήποτε γραμματική χωρίς συμφραζόμενα που δεν περιέχει έμμεση ή κρυφή αριστερή επανάληψη και δημιουργεί έναν λεκτικό αναλυτή και έναν συντακτικό αναλυτή αναδρομικής κατάβασης (recursive-descent parser) που χρησιμοποιεί προγνωστικό συντακτικό αναλυτή adaptive LL ALL(*) [112]. Τα δύο LL στην ονομασία αυτή υποδεικνύουν ότι η ανάγνωση της συμβολοσειράς εισόδου γίνεται από αριστερά

[3]`http://www.antlr.org/`

προς τα δεξιά (left-to-right) και ότι η κατασκευή του συντακτικού δέντρο αντιστοιχεί στην αριστερότερη παραγωγή (leftmost derivation). Πρόκειται για έναν τύπο συντακτικού αναλυτή από πάνω προς τα κάτω με δυνατότητα χρήσης k συμβόλων για πρόβλεψη, ο οποίος μπορεί ταυτόχρονα να λειτουργήσει προσαρμοστικά με αυθαίρετη πρόβλεψη lookahead, χωρίς σαφή καθορισμό του k. Πειράματα που παρουσιάζονται στο [112] καταδεικνύουν ότι οι αναλυτές τύπου ALL(*) παρουσιάζουν στην πράξη γραμμική συμπεριφορά, σε χρόνο και χώρο, μολονότι έχουν θεωρητική πολυπλοκότητα $O(n^4)$, και ξεπερνούν τους γενικούς αναλυτές κατά τάξεις μεγέθους.

Η προαναφερθείσα μεθοδολογία έχει υψηλή γενικότητα και δυνατότητα επέκτασης καθώς χρησιμοποιεί μια ισχυρή αφαίρεση που διαχωρίζει τον προγραμματισμό από τις γνώσεις του νομικού τομέα, ελαχιστοποιείται η πιθανότητα δημιουργίας προγραμματιστικών λαθών (bugs) και ταυτόχρονα μπορούμε εύκολα να επεκτείνουμε την γραμματική μας για να χειριστούμε περισσότερα νομικά έγγραφα π.χ. δικαστικές αποφάσεις. Παράλληλα, η υιοθέτηση μιας αρθρωτής αρχιτεκτονικής σε επίπεδα επιτρέπει την εύκολη υιοθέτηση/προσαρμογή σε νέα σχήματα/πρότυπα.

3.2.3 Διαδικασία Μοντελοποίησης Νομικών Πηγών

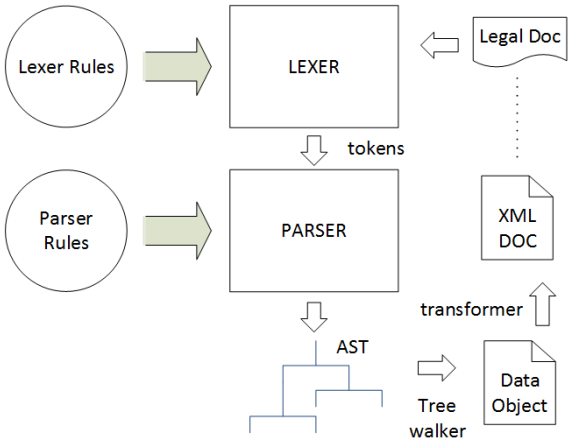
Σε αυτή την ενότητα παρουσιάζουμε λεπτομέρειες της διαδικασίας μοντελοποίησης νομικών πηγών. Στόχος μας είναι να παρέχουμε αυτόματη δομή και σημασιολογική ευρετηρίαση των νομικών πόρων. Τα βασικά βήματα της προσέγγισής μας είναι:

- Να προσδιορίσουμε τη διάρθρωση των νομικών πηγών
- Να προσδιορίσουμε τα μεταδεδομένα των νομικών πηγών
- Επικυρώσουμε τα παραγόμενα αρχεία έναντι του επιλεγμένου σχήματος

Ως στάδιο προ-επεξεργασίας, στη μέθοδο μας, όλα τα νομικά έγγραφα μετατρέπονται σε αρχεία απλού κειμένου. Ειδικότερα, αυτό το βήμα έχει το μειονέκτημα της απόρριψης πολύτιμων στοιχείων στυλ και διαρρύθμισης που απαντώνται στη μορφή του εγγράφου παρουσίασης (pdf / word). Αρκετά προγράμματα ανάλυσης περιεχομένου ανοιχτού λογισμικού μπορούν αποτελεσματικά να εκτελέσουν αυτό το έργο π.χ., Apache Tika[4]. Επίσης, η εξαγωγή κειμένου από εικόνες π.χ., σαρωμένο κείμενο αποθηκευμένο ως εικόνα στο παρεχόμενο έγγραφο μπορεί να αντιμετωπιστεί επαρκώς με μια διαδικασία OCR.

Το Σχήμα 3.3 παρέχει μια οπτική επισκόπηση της διαδικασίας αυτόματης μοντελοποίησης νομικών πηγών. Από τεχνική άποψη, το έγγραφο εισόδου μετασχηματίζεται από τον λεκτικό αναλυτή (lexer) σε ροή λεκτικών μονάδων (token), η οποία αναλύεται από τον συντακτικό αναλυτή και κατασκευάζεται το αφηρημένο συντακτικό δέντρο AST. Η διάσχιση του AST δημιουργεί ένα μοντέλο εγγράφου στην μνήμη του προγράμματος, το οποίο στη συνέχεια, ένας μετασχηματιστής μορφοποιεί στο επιλεγμένο νομικό σχήμα AKN. Τέλος, μετά την επιτυχή ολοκλήρωση του σημασιολογικού ελέγχου και επικύρωσης του παραγόμενου εγγράφου

[4] `https://tika.apache.org/`



Σχήμα 3.3: Επισκόπηση Διαδικασίας Αυτόματης Μοντελοποίησης Νομικών Πηγών

έναντι του επιλεγμένου σχήματος, παράγεται το έγγραφο εξόδου, το οποίο και αναπαριστά σε δομημένη και τυποποιημένη μορφή το νομικό έγγραφο εισόδου.

Η διαδικασία μοντελοποίησης ακολουθεί σειριακά στρατηγική αγωγού (pipeline strategy), χρησιμοποιώντας μια προσέγγιση από πάνω προς τα κάτω, η οποία μπορεί να συνοψιστεί στα ακόλουθα 5 βήματα:

1. Αναγνώριση τύπου εγγράφου. Η αναγνώριση του είδους του νομικού εγγράφου αποτελεί βασικό βήμα για την αποτελεσματική μοντελοποίηση νομικών εγγράφων. Ο τύπος του εγγράφου ορίζει τόσο τη δομή του εγγράφου, τα διαθέσιμα μεταδεδομένα όσο και την εσωτερική σημασιολογική οργάνωση του προκύπτοντος εγγράφου. Έτσι, το πρώτο βήμα στη μεθοδολογία μας είναι να προσδιορίσουμε σωστά τον τύπο του εγγράφου. Αυτό επιτυγχάνεται, σύμφωνα με τη γλώσσα ειδικού σκοπού που αναπτύχθηκε, με τη σάρωση της αρχής του κειμένου για προκαθορισμένες λέξεις-κλειδιά που διαχωρίζουν αποτελεσματικά τους τύπους εγγράφων. Η γραμματική μας αναγνωρίζει και επεξεργάζεται όλες τις νομικές πηγές που δημοσιεύονται στην Εφημερίδα της Κυβερνήσεως, δηλαδή τους νόμους, τα προεδρικά διατάγματα και τις υπουργικές αποφάσεις καθώς και διοικητικές πράξεις γενικής δομής π.χ., εγκυκλίους, αποφάσεις.

2. Δομική ανάλυση. Μετά την αναγνώριση του τύπου εγγράφου, διακρίνονται τα νομικά στοιχεία (π.χ εισαγωγικό μέρος, σώμα κειμένου, τελικό μέρος, παραρτήματα) στο έγγραφο. Η δομική/διαρθρωτική ανάλυση διαφέρει μεταξύ των διαφόρων τύπων εγγράφων όπως αυτό προκαθορίζεται στην αντίστοιχη γραμματική και εφαρμόζεται στον κατάλληλο αναλυτή.

3. Διαχωρισμός νομικών ενοτήτων - μπλοκ. Για κάθε νομικό μπλοκ που προσδιορίζεται, στα στοιχεία της δομικής ανάλυσης, γίνεται αναδρομική κατάτμηση, σύμφωνα με την αντίστοιχη γραμματική, σε ελάχιστα δομικά στοιχεία. Σε αυτό το βήμα προσπαθούμε να εντοπίσουμε και να περιγράψουμε τη δομή των νομικών εγγράφων - αυτό περιλαμβάνει τον προσδιορισμό και την περιγραφή κάθε δομικής μονάδας - νομική ενότητα του εγγράφου (τίτλος, κεφάλαια,

τμήματα, άρθρα, παραγράφους, κλπ.), ώστε αργότερα η διαδικασία προσδιορισμού και επίλυσης νομικών παραπομπών να μπορεί να προσδιορίσει και να αναφερθεί επακριβώς σε κάθε δομική μονάδα, εφόσον υπάρχουν νομικές παραπομπές προς την μονάδα αυτή. Οι τιμές μεταδεδομένων εγγράφων αναγνωρίζονται επίσης σε αυτό το βήμα. Σχεδιαστικά υποθέτουμε ότι ο παρσερ λειτουργεί χωρίς διαθέσιμα, από τον χρήστη ή το περιβάλλον εκτέλεσης, μεταδεδομένα εγγράφου και έτσι ακολουθεί μια άπληστη προσέγγιση για να εντοπίσει όσο το δυνατόν περισσότερες τιμές μεταδεδομένων.

4. Νομική Μοντελοποίηση. Σε αυτό το βήμα, ένα μοντέλο του εγγράφου κατασκευάζεται με βάση την κατάτμηση των νομικών μπλοκ και την ακολουθία διαδοχής τους. Το μοντέλο αυτό στην συνέχεια μετασχηματίζεται στο επιλεγμένο νομικό σχήμα. Ο μετασχηματισμός ακολουθεί τους κανόνες αντιστοίχησης μεταξύ όλων των νομικών μπλοκ και τα αντίστοιχα στοιχεία τους στο επιλεγμένο νομικό σχήμα (AKN). Το μοντέλο δεδομένων παρουσιάζεται στην ενότητα 3.4.2.

5. Σημασιολογικός έλεγχος και επικύρωση. Ως τελικός βήμα της διαδικασίας, ο σημασιολογικός έλεγχος ανιχνεύει τυχόν ασυνέπειες που μπορεί να περιέχει το κείμενο και τυχόν ασυμφωνίες με το επιλεγμένο σχήμα μοντελοποίησης νομικών πηγών.

3.3 Πειραματική Μελέτη

Στην ενότητα αυτή παρουσιάζουμε την αξιολόγηση της προσέγγισής μας. Η πειραματική μας μελέτη πραγματοποιήθηκε στην βάση διπλής προσέγγισης: α) ποιοτική/ποσοτική ανάλυση για την αξιολόγηση της ακρίβειας της μεθόδου μας σε σχέση με το βέλτιστο σετ και β) ανάλυση κλιμάκωσης για την αξιολόγηση της απόδοσης κατά την μεταβολή των παραμέτρων εισόδου. Περιγράφουμε πρώτα το πειραματικό σύνολο δεδομένων και στη συνέχεια προχωράμε στα πειραματικά αποτελέσματα.

3.3.1 Συλλογή Νομικών Εγγράφων

Η Ανεξάρτητη Αρχή Δημόσιων Εσόδων μας παρείχε ένα σύνολο περισσότερων από 600 νομικών εγγράφων, όπως νόμων, προεδρικών διαταγμάτων, υπουργικών αποφάσεων και κανονιστικών πράξεων, σε μορφή αδόμητου κειμένου pdf. Με βάση τα αρχεία αυτά διεξάγαμε επαναληπτικές δοκιμές, με στόχο τη αρχική μοντελοποίηση και στην συνέχεια την βελτιστοποίηση της μεθόδου μας. Για το σκοπό της μελέτης επιλέξαμε, τυχαία, 20 νόμους και κατασκευάσαμε, με την συνδρομή νομικών εμπειρογνωμόνων, την πρότυπη μοντελοποίηση τους (σύνολο αντικειμενικής αλήθειας), με βάση το σχήμα δεδομένων AKN, όπως αυτό το επεκτείναμε για να καλύψουμε τις ιδιαιτερότητες των Ελληνικών νομικών πηγών. Χρησιμοποιήσαμε το λογισμικό Apache Tika[5] για να εξάγουμε το κείμενο από τα αρχεία pdf. Τα προκύπτοντα αρχεία κειμένου δόθηκαν στην συνέχεια ως είσοδος στη μέθοδο μας, η οποία παρήγαγε αρχεία XML συμβατά με το σχήμα δεδομένων AKN.

[5] `https://tika.apache.org/`

Πίνακας 3.1: Αξιολόγηση Αποτελεσματικότητας Δόμησης Νομικών Πηγών

Method	Precision	Recall	F-measure
Structural Analysis	1.0	0.923	0.960
Legal Blocks Isolation	0.987	0.96	0.973

3.3.2 Αξιολόγηση Αποτελεσματικότητας

Η αξιολόγηση ενός συντακτικού αναλυτή [26] βασίζεται στον υπολογισμό της ακρίβειας και της ανάκλησης πάνω σε γραμματικές σχέσεις. Έτσι, αξιολογήσαμε την απόδοση του αναλυτή μας χρησιμοποιώντας τις ακόλουθες, διεθνώς καθιερωμένες, μετρικές:

– *Precision*, μετρά το ποσοστό των εξαρτήσεων με συγκεκριμένο τύπο στην έξοδο ανάλυσης που ήταν σωστά

– *Recall*, μετρά το ποσοστό των εξαρτήσεων με συγκεκριμένο τύπο στο σετ δοκιμών που έχουν αναλυθεί σωστά

– *F-measure*, ο αρμονικός μέσος όρος ακρίβειας και ανάκλησης.

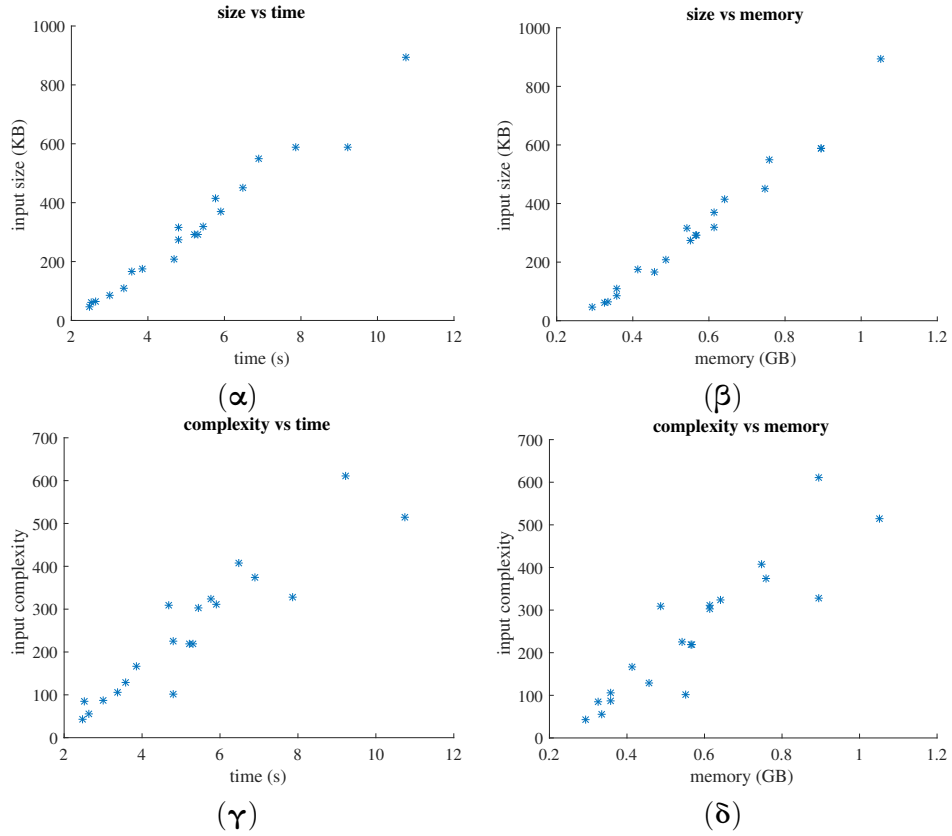
Στόχος μας είναι να αξιολογήσουμε την αποτελεσματικότητα, δηλαδή την ακρίβεια στην εξαγωγή και μοντελοποίηση νομικών πόρων από αρχεία αδόμητου κειμένου. Σημειώνουμε ότι με βάση την πολυπλοκότητα της εισόδου, μια ακριβής στατιστική ανάλυση σχετικά με την ακρίβεια/ανάκληση σε ολόκληρο το σύνολο δεδομένων (δομική ανάλυση, ανίχνευση μεταδεδομένων, uri assignment, σημασιολογική μοντελοποίηση) είναι μια απαιτητική εργασία που απαιτεί την χειροκίνητη επίσημείωση όλων νομικών μπλοκ και μεταδεδομένων των νομικών πηγών. Για το λόγο αυτό, συγκρίναμε την πρότυπη μοντελοποίηση των αρχείων με την έξοδο του parser για το δεύτερο και το τρίτο βήμα της μεθοδολογίας μοντελοποίησης (Υπό-Ενότητα 3.2.3), εξετάζοντας μόνο την αναγνώριση δομικής ανάλυσης. Τα αποτελέσματα των μετρήσεων παρουσιάζονται στον Πίνακα 3.1.

Ο αναλυτής μας παρουσιάζει υψηλά ποσοστά ακρίβειας και ανάκλησης, ποσοστά νομικών ενοτήτων που ανιχνεύονται και αναλύονται επιτυχώς. Από την άποψη της δομικής ανάλυσης (προσδιορισμός του εισαγωγικού μέρος, του σώματος κειμένου, του τελικού μέρους, των παραρτημάτων) τα σφάλματα που προέκυψαν κατά την αξιολόγηση περιελάμβαναν τον προσδιορισμό των παραρτημάτων. Στο τμήμα διαχωρισμός νομικών ενοτήτων-μπλοκ, ο συνηθέστερος τύπος σφάλματος αφορά τα σφάλματα που σχετίζονται με την ταυτοποίηση των πινάκων και τις αναντιστοιχίες στα τροποποιητικά έγγραφα (τμήματα των εγγράφων που έχουν απαριθμηθεί). Ένας άλλος, λιγότερο συχνός, τύπος σφάλματος που εντοπίσαμε ήταν η μη σωστή αναγνώριση του τίτλου παραγράφου, καθώς δεν είναι πάντα δυνατό να προσδιοριστεί εάν το κείμενο που αρχίζει στην πρώτη γραμμή ενός τμήματος αποτελεί συνέχεια ενός τίτλου ή μέρος μιας παραγράφου.

3.3.3 Ανάλυση Κλιμάκωσης

Αξιολογήσαμε την απόδοση του μηχανισμού μας σε σχέση με το μέγεθος (KB) και την νομική 'πολυπλοκότητα' (αριθμός και ποικιλία νομικών μπλοκ) των πηγών εισόδου. Το μέγεθος των πηγών εισόδου ποικίλλει από μερικά KB για τα μικρότερα αρχεία έως και $\sim 900KB$ για το μεγαλύτερο. Επίσης, η πολυπλοκότητα ποικίλει από ~ 40 έως 600 νομικές ενότητες.

Μετρήσαμε την απόδοση σε σχέση με τον χρόνο που απαιτείται για την παραγωγή της εξόδου και την κατανομή της μνήμης. Ο αναφερόμενος χρόνος αποτελεί άθροισμα του χρόνου κάθε μεμονωμένου βήματος, δηλαδή διάβασμα της εισόδου, ανάλυσης του κειμένου και αποθήκευσης της εξόδου. Αναφερόμενη μνήμη είναι η συνολική μνήμη που χρησιμοποιείται από το εκτελέσιμο αρχείο και όχι η άμεσα κατανεμημένη μνήμη στην μονάδα συντακτικής ανάλυσης.



Σχήμα 3.4: Διαγράμματα μεγέθους και δομικής πολυπλοκότητας των νομικών εγγράφων εισόδου σε σχέση με τον χρόνο που απαιτήθηκε για την ανάλυση (α,γ) και τις απαιτήσεις μνήμης (β,δ). Παρατηρούμε ότι υπάρχει μια γραμμική σχέση μεταξύ μεγέθους/πολυπλοκότητας των νομικών εγγράφων και χρόνου/μνήμης που απαιτούνται για την συντακτική τους ανάλυση.

Το Σχήμα 3.4 συνοψίζει τα αποτελέσματα της ανάλυσης επεκτασιμότητας. Στο Σχήμα 3.4 (α),(β) παρουσιάζεται η απόδοση του συστήματος, σε όρους χρόνου και απαιτήσεις μνήμης με βάση το μέγεθος εισόδου, ενώ στο Σχήμα 3.4 (γ),(δ) παρουσιάζονται τα αντίστοιχα μεγέθη με βάση την 'νομική πολυπλοκότητα' του αρχείου εισόδου. Τα περισσότερα αρχεία, που αφορούν μικρά μεγέθη και πολυπλοκότητα, μοντελοποιούνται μέσα σε λίγα δευτερόλεπτα και απαιτούν λιγότερο από 0,4 GB μνήμης. Αρχεία μεγάλου μεγέθους και αυξημένης πολυπλοκότητας

επεξεργάζονται επίσης με προσιτές απαιτήσεις χρόνου και μνήμης, με γραμμική συμπεριφορά της απόκρισης ως προς την είσοδο. Σημειώνουμε επίσης ότι αποκλίνοντα τμήματα στα Σχήματα 3.4 (γ),(δ) οφείλονται στην πολύπλοκη δομή φωλιάσματος του νομικού εγγράφου εισόδου, που αναγκάζει τον συντακτικό αναλυτή να αξιολογήσει έναν μεγάλο αριθμό κανόνων σύνταξης. Με βάση τα αποτελέσματα, μπορούμε να συμπεράνουμε ότι η υλοποίηση της μεθοδολογίας μας, στην πράξη ακολουθεί ένα γραμμικό πρότυπο όσον αφορά το μέγεθος/πολυπλοκότητα των νομικών εγγράφων και του χρόνου/μνήμης που απαιτούνται για την συντακτική τους ανάλυση.

3.4 Πλατφόρμα Διαχείρισης Νομικών Κειμένων

Σε αυτή την ενότητα αρχικά παρουσιάζουμε συνοπτικά τα βασικά χαρακτηριστικά του Solon και το μοντέλο δεδομένων που χρησιμοποιείται για την μοντελοποίηση των νομικών πηγών. Στην συνέχεια περιγράφουμε την αρχιτεκτονική της πλατφόρμας και παρουσιάζουμε μια επισκόπηση των κύριων υποσυστημάτων[6]. Τέλος, αξιολογούμε την προτεινόμενη αρχιτεκτονική, μέσω της παραγωγικής εγκατάστασης του Solon σε δημόσιο φορέα.

3.4.1 Απαιτήσεις και Γενικά Χαρακτηριστικά

Οι βασικές απαιτήσεις για το Solon, όπως προέκυψαν αρχικά από βιβλιογραφική έρευνα σχετικά με τα συστήματα διαχείρισης νομικών εγγράφων και από την ανάλυση των επιχειρηματικών απαιτήσεων για το έργο 'Σύστημα Ψηφιακής Βιβλιοθήκης (Context Sensitive Help)' (Ενότητα Πειραματική Μελέτη σε Περιβάλλον Παραγωγής 3.5), είναι:

- Υποστήριξη για αυτόματη και μη αυτόματη εισαγωγή μη δομημένων εγγράφων από προκαθορισμένες νομικές δικτυακές πήλες.
- Αυτόματη δομική ανάλυση και σημασιολογική αναπαράσταση δεδομένων και μεταδεδομένων νομικών εγγράφων.
- Αυτόματη ανακάλυψη και επίλυση νομικών παραπομπών για κάθε αντίστοιχη δομική μονάδα.
- Αυτόματη ταξινόμηση των νομικών πηγών βάσει ειδικών κανόνων.
- Υποστήριξη για τη χειρωνακτική αποκατάσταση (manual curation) του αυτόματα δομημένου και σημασιολογικά εμπλουτισμένου περιεχομένου.
- Υποστήριξη πολλαπλών κριτηρίων και πολύπλευρης αναζήτησης χρησιμοποιώντας όλα τα μεταδεδομένα που προσδιορίζονται στα έγγραφα
- Υποστήριξη για δομημένη ανάκτηση περιεχομένου.

Το Solon βασίζεται σε ένα σύνολο υποσυστημάτων που ενσωματώνονται μέσω σαφώς καθορισμένων API, επικοινωνούν μέσω διασυνδέσεων REST HTTP και μπορούν να χρησιμοποιηθούν όχι μόνο ως μέρη της συνολικής αρχιτεκτονικής αλλά και ως μεμονωμένες υπηρεσίες.

[6]Συμπληρωματικό υλικό παρέχεται στο `https://github.com/mkoniari/Solon/`

3.4.2 Μοντέλο Δεδομένων

Οι νομικές πηγές μοντελοποιούνται με βάση το σχήμα Akoma Ntoso (AKN) [14]. Το AKN αποτελεί πρότυπο OASIS, σχήμα XML για τη μοντελοποίηση κοινοβουλευτικών, νομοθετικών και δικαστικών εγγράφων[7]. Το σχήμα AKN καθιστά τις δομικές και σημασιολογικές συνιστώσες των νόμιμων πηγών πλήρως προσβάσιμες σε μηχαναγνώσιμες διαδικασίες, επιτρέποντας έτσι τη δημιουργία υψηλής ποιότητας νομοθετικών υπηρεσιών πληροφόρησης και βελτιώνοντας σημαντικά την αποτελεσματικότητα και τη λογοδοσία σε κοινοβουλευτικό, νομοθετικό και δικαστικό πλαίσιο.

Για την υποστήριξη των Ελληνικών νομικών πηγών, δηλαδή τους νόμους, τα προεδρικά διατάγματα, τις υπουργικές αποφάσεις και τις διοικητικές πράξεις, αναλύσαμε τις ελληνικές νομικές πηγές και δημιουργήσαμε κανόνες αντιστοίχησης μεταξύ όλων των νομικών μπλοκ και τα αντίστοιχα στοιχεία τους στο σχήμα AKN. Επίσης, επεκτείναμε κατάλληλα το σχήμα για να φιλοξενήσουμε προσαρμοσμένα ελληνικά μεταδεδομένα με βάση το λεξιλόγιο Dublin Core.

Συνολικά, σύμφωνα με τα πέντε επίπεδα συμμόρφωσης με το σχήμα AKN, το μοντέλο δεδομένων μας συμμορφώνεται με το επίπεδο 3. Ακολουθούμε α) τη δομή εγγράφου που ορίζεται στο σχήμα AKN (π.χ. πρόλογος, προοίμιο, σώμα κειμένου, συμπεράσματα, παραρτήματα) για ολόκληρο το έγγραφο, β) την σύμβαση ονοματοδοσίας URI/IRI (για μεταδεδομένα τύπου Functional Requirements for Bibliographic Records - FRBR[8]) και ID που ορίζονται στη Σύμβαση Ονομάτων του AKN και γ) αναγνωρίζουμε τα βασικά μεταδεδομένα FRBR, publication, normative reference.

Επιπλέον, σε ευθυγράμμιση με το πρότυπο ELI, η προσέγγισή μας προσφέρει το ελάχιστο σύνολο μεταδεδομένων που απαιτείται από την προδιαγραφή ELI και εκχωρεί ένα URI σε κάθε διαφορετικό νομικό μπλοκ. Το πρότυπο European Legislation Identifier - ELI[9] αποτελεί προτεινόμενο από την Ε.Ε. πρότυπο[10] για ένα Ευρωπαϊκό Αναγνωριστικό Νομοθεσίας που παρέχει, μεταξύ άλλων, λύση για τον μοναδικό προσδιορισμό και πρόσβαση στην εθνική και ευρωπαϊκή νομοθεσία. Με αυτό τον τρόπο η μοντελοποίηση κάθε νομικού μπλοκ συμμορφώνεται με το πρότυπο ELI, ώστε να διευκολυνθεί η ακριβής σύνδεση των νομικών παραπομπών για κάθε αντίστοιχο νομικό μπλοκ.

3.4.3 Αρχιτεκτονική

Η αρχιτεκτονική του Solon, Σχήμα 3.5, αποτελείται από διάφορα υποσυστήματα (components) που προσφέρουν τις υπηρεσίες τους στην βάση της αρχιτεκτονικής REST.

Το Αποθετήριο Νομικών Πηγών παρέχει λειτουργίες για την αποθήκευση και τη διαχείριση νομικών πηγών. Το υποσύστημα Συγκομιδής συλλέγει από απομακρυσμένες πηγές πληροφοριών έγγραφα ως δεδομένα εισόδου, τα οποία στη συνέχεια και μετατρέπονται σε μια

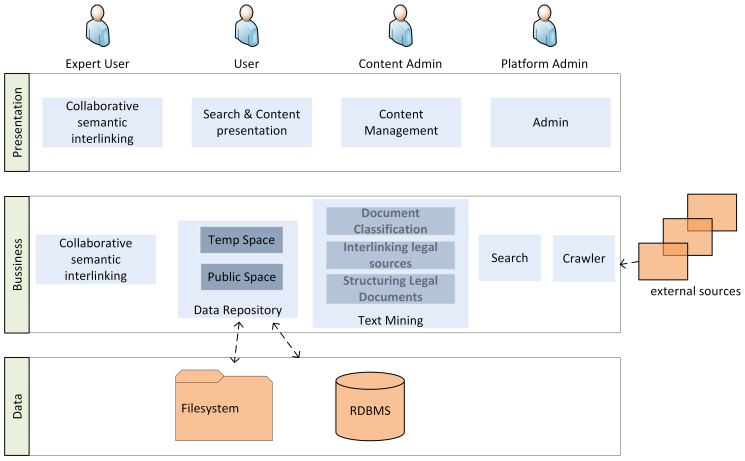
[7] `https://www.oasis-open.org/committees/legaldocml/`

[8] `https://en.wikipedia.org/wiki/Functional_Requirements_for_Bibliographic_Records`

[9] `https://en.wikipedia.org/wiki/European_Legislation_Identifier`

[10] Συμπεράσματα του Συμβουλίου για τη θέσπιση του Αναγνωριστικού Ευρωπαϊκής Νομοθεσίας (European Legislation Identifier — ELI) (2012/C 325/02)

σημασιολογικά πλούσια δομή δεδομένων στο υποσύστημα εξόρυξης κειμένου. Το υποσύστημα Αναζήτησης είναι υπεύθυνο για την αποτελεσματική ευρετηρίαση και ανάκτηση νομικών πληροφοριών. Το υποσύστημα Σημασιολογικών Παραπομπών επιτρέπει στους χρήστες να συνδέουν, να οργανώνουν και να διερευνούν τους νομικούς πόρους σύμφωνα με τις ατομικές τους προτιμήσεις/ ανάγκες. Τέλος, συμπληρωματικά με τα προαναφερθέντα λειτουργικά υποσυστήματα, το υποσύστημα διαχείρισης παρέχει υποστήριξη για τη χειρωνακτική επιδιόρθωση του περιεχομένου καθώς και υπηρεσίες διαχείρισης γενικής λειτουργικότητας.



Σχήμα 3.5: Λογική Αρχιτεκτονική Solon

3.4.4 Αποθετήριο Νομικών Πηγών

Η κύρια λειτουργία του χώρου αποθήκευσης νομικών πηγών είναι να προσφέρει μόνιμη και αξιόπιστη αρχειοθέτηση του ψηφιακού περιεχομένου με έμφαση στη διαθεσιμότητα και τη χρήση των δεδομένων, σύμφωνα με σύγχρονα πρότυπα στο διαδίκτυο.

Το αποθετήριο νομικών πηγών βασίζεται στο Flexible Extensible Digital Object Repository, Fedora[11], ένα αρθρωτό αποθετήριο, ανοιχτού λογισμικού [81]. Η αρχιτεκτονική του αποθετηρίου παρέχει τη βάση για την αποθήκευση και τη διαχείριση σύνθετων ψηφιακών αντικειμένων, καθώς και τις σχέσεις μεταξύ τους σε μοντέλο RDF. Το αποθετήριο μας, που δημιουργήθηκε με στόχο να υποβοηθήσει το σημασιολογικό επίπεδο, παρέχει δυνατότητες που διευκολύνουν τη χρήση του συστήματος και την υποδομή πρόσβασης και ανταλλαγής νομικής γνώσης.

Τα έγγραφα που αποθηκεύονται στο αποθετήριο, νομικά και διοικητικά έγγραφα, δεν ακολουθούν μόνο ένα ιεραρχικό μοντέλο δεδομένων, αλλά αποτελούνται επίσης από διάφορα έγγραφα/αρχεία, π.χ. μορφή εκτύπωσης (pdf), αναπαράσταση σε XML (σχήμα AKN), συνοδευτικές εικόνες, κλπ. Αποθηκεύονται σε ένα κατευθυνόμενο ακυκλικό γράφημα πόρων όπου οι ακμές αντιπροσωπεύουν μια σχέση γονέα-παιδιού (parent-child relation). Η διαχείριση των ψηφιακών κοντέινερ, δηλαδή οι λειτουργίες δημιουργίας, ανάγνωσης, ενημέρωσης και διαγραφής και οι λειτουργίες εισαγωγής και εξαγωγής, παρέχονται μέσω API RESTful

[11] `http://fedorarepository.org`

HTTP, ακολουθώντας την προδιαγραφή W3C Linked Data Platform[12].

3.4.5 Συγκομιδή Νομικών Πηγών

Η ανίχνευση και συγκομιδή νομικών εγγράφων από δημόσιες πύλες είναι μια σαφώς οριοθετημένη διαδικασία, που όμως εξαρτάται σε μεγάλο βαθμό από τα συγκεκριμένα χαρακτηριστικά της πηγής πληροφοριών που πρόκειται να ανιχνευθεί. Ο κύριος λόγος αυτής της ιδιαιτερότητας είναι η έλλειψη ομοιομορφίας όσον αφορά την αλληλεπίδραση με τον πάροχο δεδομένων ή το πραγματικό περιεχόμενο που παρέχεται. Οι νομικές πηγές συνήθως παρέχονται με αδόμητο τρόπο, μέσω διαφόρων μορφών. Συνοδευτικά μεταδεδομένα μπορεί να υπάρχουν ή όχι, σε διάφορες μορφές. Επιπλέον, ένα ενιαίο νομικό έγγραφο μπορεί να αποτελείται από διάφορα χωριστά προσφερόμενα έγγραφα, τα οποία πρέπει να μεταφορτωθούν μαζί και τελικά να ενωθούν, π.χ., αναπαράσταση κειμένων και συνοδευτικές εικόνες. Τέλος, τα νομικά έγγραφα ενδέχεται να μην είναι άμεσα προσβάσιμα μέσω μιας κλήσης API, αλλά να προσφέρονται μέσω αποτελεσμάτων αναζήτησης, διαφορετικής δομής για κάθε πάροχο πληροφοριών, που το υποσύστημα συγκομιδής θα πρέπει να αναλύσει και να αξιολογήσει για να ανακτήσει το σχετικό περιεχόμενο.

Με βάση τις προαναφερθείσες πτυχές, με στόχο επίσης τη διασφάλιση συμβατότητας και κεντρικής αποθήκευσης δεδομένων, επιλέχθηκε μια κατανεμημένη αρχιτεκτονική για το σχεδιασμό του υποσυστήματος συγκομιδής. Αποτελείται από α) έναν διαχειριστή συγκομιδής και β) από διάφορες εφαρμογές συγκομιδής, οι οποίες επεκτείνουν μια κοινή διεπαφή 'συγκομιδής', αλληλεπιδρώντας με τον διαχειριστή. Ο διαχειριστής συγκομιδής είναι κυρίως υπεύθυνος για την αλληλεπίδραση με το αποθετήριο νομικών πηγών, εξασφαλίζοντας επικύρωση και αποθήκευση των δεδομένων που αποκτήθηκαν, με ιδιαίτερη μέριμνα για τον χειρισμό διπλοτύπων. Είναι επίσης υπεύθυνος για τον περιοδικό προγραμματισμό και την παρακολούθηση της απόδοσης όλων των δραστηριοτήτων συγκομιδής στην πλατφόρμα, εξασφαλίζοντας ομαλή απόδοση. Κάθε ξεχωριστή εφαρμογή συγκομιδής, με βάση τα μοναδικά χαρακτηριστικά του παρόχου πληροφοριών που πρόκειται να χρησιμοποιηθεί, είναι υπεύθυνη για την εύρεση και λήψη νέων ή επικαιροποιημένων δεδομένων και την παράδοση των σχετικών δεδομένων στο διαχειριστή συγκομιδής. Δεδομένου ότι, οι εφαρμογές συγκομιδής αλληλεπιδρούν με συγκεκριμένους παρόχους πληροφοριών, ανεξάρτητα μεταξύ τους, μπορεί να βρίσκονται στην ίδια ή σε διαφορετικές εικονικές μηχανές (VMs), παρέχοντας έτσι δυνατότητες εξισορρόπησης του φόρτου του συστήματος (load balancing).

Με την επιτυχή επικύρωση, από τον διαχειριστή συγκομιδής, των δεδομένων που είναι διαθέσιμα στο σύστημα, τα δεδομένα στη συνέχεια αποθηκεύονται σε έναν προσωρινό χώρο εργασίας του αποθετηρίου, ξεκινώντας τις διαδικασίες εξόρυξης κειμένου.

3.4.6 Εξόρυξη κειμένου

Η διαδικασία εξόρυξης κειμένου συνίσταται στην σειριακή εφαρμογή, με στρατηγική αγωγού (pipeline strategy, διαδοχικών μετασχηματισμών εμπλουτισμού των νομικών πηγών.

[12] `http://www.w3.org/TR/ldp/`

Οι κύριες εργασίες που επιτελούνται αφορούν την διαδικασία μοντελοποίησης νομικών πηγών, τη διασύνδεση των νόμιμων πηγών και την ταξινόμηση τους με χρήση ειδικών κανόνων.

Διαδικασία Μοντελοποίησης Νομικών Πηγών Η εν λόγω διαδικασία αναλύθηκε διεξοδικά στην Ενότητα 'Αυτόματη Μοντελοποίηση και Σημασιολογική Ανάλυση Νομικών Κειμένων' (3.2).

Διαδικασία Διασύνδεσης Νομικών Πηγών. Απαραίτητες προϋποθέσεις για την αποτελεσματική διάδοση των νομικών πηγών είναι τόσο η ηλεκτρονική διαθεσιμότητα των νομικών πηγών σε δομημένο και τυποποιημένο μορφότυπο όσο και η διασύνδεση τους, εγκαθιστώντας διασυνδέσεις εντός του ίδιου ή παρόμοιων αποθετηρίων. Δεδομένης της άφθονης χρήσης παραπομπών μεταξύ των νομικών πηγών, η χειρωνακτική επισήμανση είναι αρκετά δαπανηρή, τόσο από πλευράς χρόνου όσο και από πλευράς προσπάθειας που απαιτείται, καθιστώντας την ως λιγότερο ελκυστική επιλογή. Οι μέθοδοι για την αυτόματη εξαγωγή των νομικών αναφορών και την διασύνδεση των πηγών παρέχουν μια βιώσιμη εναλλακτική λύση.

Αν και οι νομικές παραπομπές θα πρέπει θεωρητικά να ακολουθούν μια προβλέψιμη δομή, καθώς οι επίσημες κατευθυντήριες γραμμές για τη νομοθετική διατύπωση θεσπίζουν κανόνες για την σύνταξη των νομικών παραπομπών, οι αποκλίσεις από τις επίσημες κατευθυντήριες γραμμές αποτελούν τον κανόνα και όχι την εξαίρεση. Το φαινόμενο αυτό αποδίδεται σε πολλούς διαφορετικούς παράγοντες, π.χ., αλλαγές στα πρότυπα αναφοράς με την πάροδο του χρόνου, ανθρώπινος παράγοντας κλπ.

Το έργο της ανίχνευσης και της επίλυσης των νομικών αναφορών συνίσταται σε διάφορα στάδια. Αρχικά, ακολουθώντας την ορολογία που καθορίστηκε στο [35], σε συνεργασία με νομικούς εμπειρογνώμονες προσδιορίσαμε και κατηγοριοποιήσαμε τους τύπους αναφορών, π.χ. απλές, σύνθετες, πλήρης και ατελής και κατασκευάσαμε σύνολα συγκεκριμένων προτύπων συμβολοσειρών. Με βάση αυτά, στην συνέχεια δημιουργήσαμε σύνολα κανονικών εκφράσεων και αναπτύξαμε αλγορίθμους για τον προσδιορισμό των νομικών παραπομπών. Η έρευνά μας αρχικά επικεντρώθηκε σε αναφορές μεταξύ των ελληνικών νομικών πηγών, αλλά στη συνέχεια επεκτείναμε τον τομέα εφαρμογής για να καλύψουμε τις αναφορές από ελληνικές σε κοινοτικές νομικές πηγές.

Μετά την 'ανακάλυψη' μιας αναφοράς, ακολουθεί η επίλυση (resolution) της αναφοράς, προσδιορίζοντας τα URI των πηγών που αναφέρονται. Όπως αναφέρθηκε ήδη, η διαδικασία μοντελοποίησης νομικών πηγών παράγει αναγνωριστικά για κάθε δομική μονάδα, σύμφωνα με το πρότυπο ELI. Ο ανιχνευτής συνδέσεων, ακολουθεί τις ίδιες αρχές και κατασκευάζει URI συμβατά με το πρότυπο ELI για κάθε αναφορά που ανακαλύφθηκε. Για να αντιμετωπίσουμε τις περιπτώσεις ονοματικών προσδιορισμών (aliases) των νομικών πηγών, όπως 'Αστικός Κώδικας', οι οποίες δεν μπορούν να επεξεργαστούν αυτόματα, έχουμε συντάξει μια βάση δεδομένων σχετικών πληροφοριών, η οποία λειτουργεί ως μηχανισμό επίλυσης, ανάμεσα σε γνωστούς ονοματικούς προσδιορισμούς στον τομέα του δικαίου και τα αναγνωριστικά όσον αφορά το URI.

Κατηγοριοποίηση Νομικών Πηγών Ο μηχανισμός κατηγοριοποίησης του Solon βασίζεται σε μια εξειδικευμένη μηχανή κανόνων που ακολουθεί ντετερμινιστική προσέγγιση. Οι κανόνες ορίζονται μέσω του διαχειριστικού τμήματος γραφικής διαπροσωπείας χρήστη και εκτελούνται έναντι των νομικών πηγών χρησιμοποιώντας προτεραιότητες. Οι κανόνες μπορούν να είναι απλοί ή να συνδυαστούν μεταξύ τους, σχηματίζοντας σύνθετους κανόνες. Ενεργούν στην βάση των μοντελοποιημένων νομικών πηγών ή των μεταδεδομένων τους, για παράδειγμα 'εάν η υπηρεσία έκδοσης της διοικητικής πράξης είναι x, ο υπογράφων του εγγράφου είναι y και η ημερομηνία έκδοσης βρίσκεται εντός του εύρους z ταξινόμησε την νομική πηγή ως w'. Επιπρόσθετα, επί του παρόντος, εξετάζουμε τεχνικές μηχανικής μάθησης για την αυτόματη ταξινόμηση των νομικών πηγών με περιγραφές από το EuroVoc[13], ένα πολυγλωσσικό, πολυεπιστημονικό θησαυρό που καλύπτει τις δραστηριότητες της ΕΕ, και χρησιμοποιείται, μεταξύ άλλων, για την επισημείωση των νομικών πηγών της Ε.Ε..

3.4.7 Σημασιολογικές Παραπομπές

Η κεντρική φιλοσοφία του Solon είναι να καταστήσει ευκολότερη την πρόσβαση σε νομικές πηγές/πληροφορίες. Επικουρικά προς την κατεύθυνση αυτή συνδράμει και η υιοθέτηση ενός τρόπου οργάνωσης της νομικής πληροφορίες ανάλογα την οπτική γωνία έκαστου χρήστη και η δυνατότητα διαμοιρασμού αυτής της οπτικής γωνίας με άλλους χρήστες. Οι σχολιασμοί (annotations) είναι ένας δημοφιλής μηχανισμός για την ενσωμάτωση της γνώσης του τελικού χρήστη στις διαδικασίες ψηφιακής επεξεργασίας και την αύξηση των ψηφιακών στοιχείων με πρόσθετες πληροφορίες. Σε προηγούμενη εργασία [96] παρουσιάστηκε η υποδομή που επιτρέπει στους χρήστες να επαναχρησιμοποιήσουν καθιερωμένες οντολογίες καθώς και σημασιολογία που δημιουργούν για να σχολιάσουν και να μοιραστούν πόρους πάνω σε δυναμικά μεταβαλλόμενα περιβάλλοντα χρήσης. Το υποσύστημα σημασιολογικών παραπομπών στοχεύει στην ενίσχυση ενός συνεργατικού περιβάλλοντος για τη διαχείριση νομικών πόρων με γνώση του νομικού τομέα. Οι χρήστες μπορούν να αξιοποιήσουν τις τεχνολογίες συνδεδεμένων δεδομένων (linked data technologies) για τη δημιουργία, οργάνωση και αξιοποίηση συνεργατικών χώρων πληροφοριών που περιέχουν διάφορες νομικές πηγές που φιλοξενούνται στην πλατφόρμα. Με τον τρόπο αυτό παρέχεται ένα στρώμα σημασιολογίας, το οποίο μοιράζεται εγγενώς μεταξύ των τελικών χρηστών, βοηθώντας τους να διαχειρίζονται, να συνδέουν, να οργανώνουν και να διερευνούν τους νομικούς πόρους με διάφορους τρόπους. Επιπλέον, προσφέρει έναν διαισθητικό τρόπο αναζήτησης νομικών πόρων και διερεύνησης μέσω μιας faceted λειτουργίας περιήγησης.

3.4.8 Αναζήτηση

Ο στόχος κάθε συστήματος ανάκτησης πληροφοριών είναι η παροχή περιεχομένου που να ταιριάζει με ακρίβεια και να ικανοποιεί τις ανάγκες των χρηστών. Εφόσον η δυνητική ομάδα χρηστών περιλαμβάνει ετερόκλητο πλήθος ατόμων με διαφορετικό υπόβαθρο και εμπειρογνωμοσύνη, από νομικούς εμπειρογνώμονες σε απλούς χρήστες, χρησιμοποιήσαμε διάφορες

[13] `http://eurovoc.europa.eu`

τεχνικές, εκμεταλλευόμενοι διάφορες διαστάσεις/χαρακτηριστικά των νομικών πηγών, ώστε να ικανοποιήσουμε με ακρίβεια τις πληροφοριακές απαιτήσεις των χρηστών. Το υποσύστημα ανάκτησης πληροφοριών υλοποιήθηκε με βάση το solr[14], μια δημοφιλή μηχανή αναζήτησης ανοιχτού λογισμικού, και προσφέρει υπηρεσίες αναζήτησης πλήρους κειμένου, ευπροσάρμοστη αναζήτηση, ευρετηρίαση σε πραγματικό χρόνο, υποστήριξη προηγμένης προσαρμογής μέσω μιας επεκτάσιμης αρχιτεκτονικής σε συνεργασία με το αποθετήριο νομικών πηγών μας.

Στις παραδοσιακές τεχνικές ανάκτησης πληροφορίας [60] τα έγγραφα ευρετηριοποιούνται και ανακτώνται ως μεμονωμένες ατομικές μονάδες. Ωστόσο, η φύση των νομικών πηγών, π.χ. τα νομικά έγγραφα, τείνουν να έχουν αρκετά μεγάλο μέγεθος, καλύπτουν πολλαπλά θέματα, έχουν ιεραρχική δομή ενθέτων στοιχείων, συνεπάγεται ότι απαιτείται μια πιο εξειδικευμένη προσέγγιση. Ως εκ τούτου, ακολουθούμε τεχνικές δομημένης ανάκτησης [8], επιτρέποντας τον συνδυασμό κειμενικών κριτηρίων, φράσεων φυσικής γλώσσας, με διαρθρωτικά κριτήρια, δηλαδή περιορισμούς ως προς τις μονάδες αναζήτησης. Με την εξέταση των νομικών πηγών ως αθροισμάτων αλληλένδετων δομικών στοιχείων που ευρετηριάζονται και ανακτώνται, τόσο στο σύνολό τους όσο και ξεχωριστά, σύμφωνα με τις απαιτήσεις των χρηστών, μπορούμε να εκτελούμε πολύπλοκα ερωτήματα που συνδυάζουν κατηγορήματα (predicates) μεταδεδομένων και κειμένου σε ένα ενιαίο ερώτημα. Με αυτόν τον τρόπο η προσέγγισή μας προσφέρει πιο ακριβή και συναφή αποτελέσματα, ωστόσο οδηγεί σε περιττές ευρετηριάσεις, καθώς το κείμενο που εμφανίζεται σε βάθος i του δέντρου εγγράφων είναι τελικά ευρετηριασμένο i φορές.

Επιπλέον, χρησιμοποιούμε μεθόδους διαφοροποιημένης ανάκτησης αποτελεσμάτων αναζήτησης, ως μέσο βελτίωσης της ικανοποίησης των χρηστών, αυξάνοντας την ποικιλία των πληροφοριών που εμφανίζονται στον χρήστη. Σε προηγούμενες εργασίες μας, [72, 73, 74], που παρουσιάζονται στο Κεφάλαιο 4, αρχικά εισάγαμε την έννοια της διαφοροποίησης των αποτελεσμάτων αναζήτησης νομικής πληροφορίας, αναλύσαμε τον αντίκτυπο των διαφόρων χαρακτηριστικών των νομικών πηγών στον υπολογισμό των τιμών ομοιότητας ερωτήματος-εγγράφου και εγγράφου-εγγράφου, προτείναμε ειδικά κριτήρια διαφοροποίησης για το συγκεκριμένο τομέα και διεξήγαγε μια εξαντλητική αξιολόγηση διαφόρων δημοφιλών μεθόδων από τις περιοχές διαφοροποίησης αποτελεσμάτων αναζήτησης, ανάκτησης πληροφορίας σε γράφους και περίληψης κειμένων. Στο υποσύστημα ανάκτησης πληροφοριών του Solon, παράλληλα με την προεπιλεγμένη διαδικασία κατάταξης, προσφέρουμε εναλλακτικά μοντέλα κατάταξης βασισμένα σε τεχνικές διαφοροποίησης, επιδεικνύοντας αξιοσημείωτες βελτιώσεις όσον αφορά τον εμπλουτισμό των αποτελεσμάτων αναζήτησης με αλλιώς κρυφές πτυχές του νομικού χώρου του ερωτήματος.

3.5 Πειραματική Μελέτη σε Περιβάλλον Παραγωγής

Μέρος της αρχιτεκτονικής του Solon υλοποιήθηκε προσφάτως, καλύπτοντας τις απαιτήσεις του έργου 'Σύστημα Ψηφιακής Βιβλιοθήκης (Context Sensitive Help)'. Η πλατφόρμα λειτουργεί σε περιβάλλον παραγωγής [15], υπό την εποπτεία της Ανεξάρτητης Αρχής Δημόσιων

[14] `http://lucene.apache.org/`
[15] `http://www.publicrevenue.gr/elib/`

Εσόδων - ΑΑΔΕ[16], με σκοπό την παροχή σημασιολογικής πρόσβασης στην ελληνική φορολογική νομοθεσία. Αυτή τη στιγμή φιλοξενεί περισσότερα από 4000 νομικά και διοικητικά έγγραφα. Ο σκοπός του έργου από την άποψη της ΑΑΔΕ είναι 'να προσφέρει στον πολίτη, αλλά και στους εσωτερικούς χρήστες, ένα εύχρηστο, ηλεκτρονικό εργαλείο εντοπισμού και ανάκτησης εγγράφων νομικού και κανονιστικού περιεχομένου, με τρόπο εύκολο, ακριβή και επίκαιρο. Επιπλέον, μέσω της Ηλεκτρονικής Βιβλιοθήκης, θα παρέχονται πληροφορίες εντοπισμένες στο ακριβές πλαίσιο εννοιολογικής αναφοράς, χωρίς περιττή πληροφορία' [119].

Στην συγκεκριμένη μελέτη περίπτωσης μας απασχολεί ιδιαίτερα η αυθεντικότητα των παρόχων πληροφορίας. Στις πρωτογενείς και δευτερογενείς πηγές δικαίου, όπως οι νόμοι, οι κανονισμοί και οι διοικητικές πράξεις, η κυβέρνηση είναι ο συγγραφέας και κυβερνητικοί ιστότοποι είναι εξουσιοδοτημένοι εκδότες. Υλοποιήσαμε τρεις εξωτερικές πηγές νομικών πληροφοριών: α) Επίσημη Εφημερίδα της Ελληνικής Δημοκρατίας[17], η οποία αναρτά νόμους, προεδρικά διατάγματα και υπουργικές αποφάσεις, β) Διαδικτυακή Πύλη Διαύγεια [18], όπου όλοι οι φορείς της ελληνικής κυβέρνησης υποχρεούνται να μεταφορτώσουν τις πράξεις και τις αποφάσεις τους και γ) ως εναλλακτική πηγή δεδομένων, χρησιμοποιούμε ειδικά αιτήματα μεταφόρτωσης περιεχομένου, που εκτελούνται από εξουσιοδοτημένους χρήστες μέσω κατάλληλης διεπαφής.

Στη συνέχεια, οι μη δομημένες νομικές πηγές μετασχηματίζονται από τον αναλυτή σε μια μηχαναγνώσιμη, σημασιολογική αναπαράσταση, η οποία και εμπλουτίζεται από τις διαδικασίες διασύνδεσης νομικών πηγών και ταξινόμησης. Η διαδικασία χειρωνακτικής επιδιόρθωσης, που εκτελείται από δημόσιους υπαλλήλους, μέσα από κατάλληλη διαχειριστική διεπαφή, ρυθμίζει την αυτόματη διαδικασία και διορθώνει τυχόν αστοχίες/ατέλειες. Οι προηγμένες υπηρεσίες αναζήτησης που προσφέρονται από το υποσύστημα αναζήτησης παρέχουν ακριβή και ποικιλόμορφα αποτελέσματα. Τα αποτελέσματα αυτά, με βάση την παρεχόμενη λειτουργικότητα του υποσυστήματος σημασιολογικών παραπομπών οι χρήστες έχουν την δυνατότητα να τα οργανώσουν σύμφωνα με τις ατομικές τους ανάγκες και να εξερευνήσουν τους νομικούς πόρους με διάφορους τρόπους.

Οι τελικοί χρήστες, δημόσιοι υπάλληλοι, εμπειρογνώμονες και πολίτες, εντυπωσιάστηκαν θετικά με τη χρήση της πλατφόρμας. Πραγματοποιήσαμε αρκετές συνεντεύξεις με εκπροσώπους από τις προαναφερόμενες κατηγορίες τελικών χρηστών, συλλέγοντας τα σχόλιά τους και λάβαμε αρκετές προτάσεις, οι οποίες εστιάστηκαν κυρίως στη βελτίωση της χρηστικότητας της εφαρμογής. Με βάση την αρχική ανατροφοδότηση, προετοιμάζουμε μια on line έρευνα για την ποσοτική και ποιοτική αξιολόγηση των παρεχόμενων υπηρεσιών, όπως την εκλαμβάνουν οι τελικοί χρήστες. Μεμονωμένες μέθοδοι και στοιχεία της πλατφόρμας μας έχουν αξιολογηθεί εκτενώς [78, 73, 74], αλλά σχεδιάζουμε επίσης μια ολοκληρωμένη τεχνική αξιολόγηση της πλατφόρμας συνολικά.

[16] `http://www.aade.gr/`, πρωην Γενική Γραμματεία Δημόσιων Εσόδων - ΓΓΔΕ
[17] `http://www.et.gr`
[18] `https://diavgeia.gov.gr/en`

3.6 Συμπεράσματα

Σε αυτό το κεφάλαιο παρουσιάσαμε το Solon μια αρχιτεκτονική κατάλληλη για τη μοντελοποίηση, τη διαχείριση και την εξόρυξη νομικών πηγών. Χρησιμοποιεί μια πρότυπη μέθοδο για την εξαγωγή σημασιολογικών αναπαραστάσεων νομικών πηγών από μη δομημένες μορφές, τη διασύνδεση και ταξινόμηση τους. Παρέχει επίσης εξειδικευμένα και διαφοροποιημένα αποτελέσματα αναζήτησης χρησιμοποιώντας τη δομή και τα ειδικά χαρακτηριστικά των νόμιμων πηγών, ενώ επιτρέπει στους χρήστες να συνδέουν, να οργανώνουν και να διερευνούν τους νομικούς πόρους ανάλογα με τις ατομικές τους ανάγκες.

Ως αποτέλεσμα, η ελληνική νομική γνώση παρέχεται με ανοικτά μηχαναγνώσιμα πρότυπα, εμπλουτισμένη με μεταδεδομένα που επιτρέπουν την σημασιολογική διαχείρισή, καθώς και την περαιτέρω αξιοποίησή από άλλα συστήματα και εφαρμογές, συμβάλλοντας στον εξορθολογισμό και την συστηματοποίηση της νομοπαραγωγικής διαδικασίας και της διαδικασίας παραγωγής κανονιστικών πράξεων παράλληλα με τα ευρωπαϊκά και διεθνή πρότυπα νομοθέτησης και νομολογίας.

Επεκτάσεις αυτής της έρευνας βρίσκονται υπό διερεύνηση. Αυτές περιλαμβάνουν την εφαρμογή τεχνικών επεξεργασίας φυσικής γλώσσας για τον προσδιορισμό των ονοματικών οντοτήτων που κατονομάζονται στα νομικά κείμενα και την χρονική διαχείριση των νομικών πόρων (π.χ. προσδιορισμός της χρονικής ισχύος ενός νομικού μπλοκ, διάρκεια ισχύος νομικών παραπομπών), την αυτόματη κωδικοποίηση της (προτεινόμενης) νομοθεσίας βάσει του πρωτοτύπου και των τροποποιητικών του εγγράφων (soft encoding) και, τέλος, την επέκταση της γλώσσας μοντελοποίησης νομικών πηγών για την κάλυψη δικαστικών αποφάσεων.

Κεφάλαιο 4

Δίκτυο Νομοθεσίας: Μοντελοποίηση και Ανάλυση του Δικαίου της Ε.Ε.

Οι νομοθέτες, οι σχεδιαστές νομικών συστημάτων πληροφοριών, καθώς και οι πολίτες αντιμετωπίζουν συχνά προβλήματα λόγω της αλληλεξάρτησης των νόμων και του αριθμού των αναφορών που απαιτούνται για την ερμηνεία τους. Στο κεφάλαιο αυτό παρουσιάζουμε ένα μοντέλο αναπαράστασης του δικαίου, το 'Δίκτυο Νομοθεσίας', ως μια νέα προσέγγιση για την αντιμετώπιση των προκλήσεων όσον αφορά τον εντοπισμό και την ποσοτικοποίηση της πολυπλοκότητας στο νομικό γίγνεσθαι. Το Δίκτυο Νομοθεσίας είναι ένα δίκτυο πολλαπλών σχέσεων που φιλοξενεί την ιεραρχία μεταξύ των πηγών του δικαίου και μπορεί να αντιπροσωπεύει σχέσεις διαφόρων κατηγοριών μεταξύ των νομικών πηγών, μαζί με τη χρονική εξέλιξή τους, προσφέροντας μια συστηματική εναλλακτική δομή σε ένα φυσικά εξελισσόμενο κανονιστικό σύστημα [75, 71].

4.1 Κίνητρο και Συνεισφορά

Τα σύνθετα δίκτυα (complex networks) είναι ένας νέος τομέας της επιστημονικής έρευνας εμπνευσμένος από την εμπειρική μελέτη δικτύων του πραγματικού κόσμου, όπως τα δίκτυα υπολογιστών, τα κοινωνικά και βιολογικά δίκτυα. Η μοντελοποίηση πολύπλοκων συστημάτων μέσω σύνθετων δικτύων βοηθά στην καλύτερη κατανόησή τους και δίνει την δυνατότητα της προσομοίωσης και ανάλυσής τους, καθώς και της ανάλυσης και πρόβλεψης της συμπεριφοράς διεργασιών που πραγματοποιούνται σε αυτά, όπως ο εντοπισμός, αφανών οργανωτικών αρχών, η ερμηνεία της επίδρασης της δομής του δικτύου σε μεμονωμένους κόμβους και η ποσοτικοποίηση της σχετικής σημασίας ενός κόμβου στο δίκτυο. Η έρευνα έχει δείξει ότι τα δίκτυα σε κοινωνικά, φυσικά και τεχνολογικά συστήματα δεν είναι τυχαία, αλλά ακολουθούν μια σειρά από βασικές αρχές οργάνωσης στη δομή και την εξέλιξη τους [103], διαχωρίζοντας τα έτσι από τυχαία συνδεδεμένα δίκτυα [38]. Η κατανόηση των στατιστικών ιδιοτήτων των δικτύων, μπορεί να μας δώσει απαντήσεις στις καίριες ερωτήσεις του πώς δημιουργούνται τα

δίκτυα και ποια είναι η γενικότερη συμπεριφορά τους, αντιμετωπίζοντας τα ως μία οντότητα και όχι κοιτώντας αποσπασματικά συγκεκριμένους κόμβους ή ακμές. Ο χαρακτηρισμός της τοπολογίας ενός δικτύου, οι τρόποι με τους οποίους οι διάφορες δομές δικτύου επηρεάζουν την συμπεριφορά των μοντελοποιημένων υποκειμένων, οι πιθανές δομές δικτύων που μπορεί να προκύψουν σε ένα σύστημα, η αναζήτηση πληροφορίας σε δομές δικτύων αποτελούν κάποια από τα ανοιχτά ερευνητικά ζητήματα στον τομέα αυτό.

Το γραπτό δίκαιο είναι μια συλλογή από διαφορετικά κανονιστικά έγγραφα, η οποία εξελίσσεται με την πάροδο του χρόνου, καθώς καινούργια έγγραφα συνεχώς δημιουργούνται και υφιστάμενα τροποποιούνται η ακυρώνονται. Ταυτόχρονα ένας κανόνας δικαίου συνήθως δε στέκεται μόνος του αλλά εντάσσεται σε ένα ευρύτερο σύστημα ρύθμισης, το οποίο διέπεται από κάποιες γενικότερες αρχές. Ο κάθε κανόνας θα πρέπει επομένως να ερμηνευθεί με βάση τις αρχές που διέπουν όλο το σύστημα ρυθμίσεων στο οποίο είναι ενταγμένος. Στο πλαίσιο αυτό, ο βαθμός δυσκολίας εύρεσης, κατανόησης, αλλά και συστηματικής ερμηνείας του ρυθμιστικού πλαισίου αυξάνει τόσο για τους πολίτες, όσο και για τους ειδικούς του χώρου. Ταυτόχρονα η διαδικασία κατάρτισης συνεκτικού και συνεπούς νομοθετικού πλαισίου καθίσταται όλο και πιο δύσκολο έργο. Η δημιουργία καινούργιας ή η τροποποίηση της υφιστάμενης νομοθεσίας είναι περίπλοκες διαδικασίες. Ως αποτέλεσμα, οι αρχές σε ευρωπαϊκό, εθνικό και τοπικό επίπεδο εξετάζουν συχνά τους προτεινόμενους κανονισμούς για μήνες ή χρόνια πριν τεθούν σε ισχύ. Επομένως, είναι κρίσιμο να εργαστούμε για την παροχή ενός μοντέλου που θα μας βοηθήσει να αποκαλύψουμε τις αναδυόμενες εξαρτήσεις μεταξύ του σώματος της νομοθεσία.

Προς την κατεύθυνση αυτή

1. Προτείνουμε μία πρωτότυπη προσέγγιση για την μοντελοποίηση του δικαίου σε νομικά συστήματα τύπου (civil law), όπως το Ελληνικό σύστημα δικαίου. Εφαρμόσαμε το μοντέλο που περιγράψαμε για την μοντελοποίηση του δικαίου της Ε.Ε., συλλέγοντας και αναλύοντας ένα εκτεταμένο σύνολο δεδομένων που καλύπτει το σύνολο του δικαίου της Ε.Ε., όπως αυτό δημοσιεύθηκε στην Επίσημη Εφημερίδα της Ευρωπαϊκής Ένωσης, σε χρονικό διάστημα άνω των 60 ετών.

2. Μελετήσαμε την δομή και τα τοπολογικά χαρακτηριστικά του δικτύου νομοθεσίας και καταλήξαμε στο συμπέρασμα ότι είναι ένα δίκτυο τύπου 'νόμου δύναμης μικρού κόσμου' (power-law small world network). Η δομή του νόμου δύναμης (power-law) σημαίνει ότι μερικές νομικές πηγές είναι ιδιαίτερα συνδεδεμένες και έχουν μεγάλη επιρροή, ενώ η πλειοψηφία μοιράζεται μόνο μερικές συνδέσεις. Το φαινόμενο μικρού κόσμου απαντάται σε συστήματα που θεωρούνται ιδιαιτέρως αποδοτικά, τόσο σε καθολικό, όσο και σε τοπικό επίπεδο, όσον αφορά την δυναμική των διεργασιών που συντελούνται σε αυτά π.χ., του πόσο αποτελεσματικά ανταλλάσσονται πληροφορίες μέσω του δικτύου. Φανερώνει ότι οι νόμοι συχνά συνδέονται με τους γειτονικούς νόμους, δημιουργώντας συστάδες, οι οποίες εμφανίζονται επίσης σε μεγαλύτερες κλίμακες. Με τον τρόπο αυτό είναι εφικτή η μετακίνηση από το ένα τμήμα του δικτύου σε ένα άλλο σε ένα μικρό αριθμό βημάτων. Η επικοινωνία μεταξύ των πολύ συσσωρευμένων περιοχών, αραιά συνδεδεμένων κόμβων, διατηρείται από μερικούς κόμβους, καθώς το Δίκτυο Νομοθεσίας είναι επίσης εξαιρετικά

ετερογενές σε σχέση με τον αριθμό των συνδέσεων των νομικών πηγών. Η προέλευση αυτής της ετερογένειας, μπορεί να εξηγηθεί από τη διαδικασία της επιλεκτικής προσκόλλησης, η οποία ενισχύει τη δημοτικότητα των πηγών υψηλής κατάταξης.

3. Δεδομένου ότι η νομοθεσία εξελίσσεται με την πάροδο του χρόνου, αξιολογήθηκε η χρονική εξέλιξη της νομοθεσίας της Ε.Ε. Κυριότερα συμπεράσματα είναι ότι η ισχύουσα νομοθεσία αυξάνει με την πάροδο του χρόνου σε σχέση με όλους τους τύπους τομέων και κατηγοριών αναφορών και επιπρόσθετα το Δίκτυο Νομοθεσίας ακολουθεί το νόμο πύκνωσης, δηλαδή ο αριθμός των ενεργών συσχετίσεων μεταξύ των νομικών εγγράφων αυξάνεται ταχύτερα από τον αριθμό των ενεργών νομικών εγγράφων.

4. Πραγματοποιήσαμε μια πρωτότυπη αξιολόγηση σταθερότητας της νομοθεσίας με στόχο την κατανόηση και πρόβλεψη της συμπεριφοράς του συστήματος σε περιπτώσεις αλλαγών. Τα αποτελέσματα της πειραματικής αξιολόγησης ακολουθούν την βιβλιογραφία περί δικτύων τύπου νόμου δύναμης μικρού-κόσμου καθώς το Δίκτυο Νομοθεσίας εμφανίζει ανθεκτικότητα σε τυχαίες διαταραχές και είναι ιδιαίτερα ευάλωτο σε στοχευμένη επίθεση κατά των υψηλά συνδεδεμένων κόμβων του.

4.2 Μοντελοποίηση του δικαίου της Ε.Ε.

Ο τρόπος με τον οποίο οι νομικές πηγές συσχετίζονται οδηγεί σε μια φυσική αναπαράσταση του γραπτού δικαίου, το δίκτυο. Εάν θεωρήσουμε το σύνολο των νομικών πηγών ως κόμβους και τις μεταξύ τους αναφορές ως τις ακμές, έχουμε ένα κατευθυνόμενο δίκτυο. Ωστόσο, προκειμένου να μοντελοποιήσουμε πλήρως το γραπτό δίκαιο θα πρέπει να το αναλύσουμε διεξοδικά και προσεκτικά να προσδιορίσουμε τα μοναδικά χαρακτηριστικά του. Στις ακόλουθες υπό-ενότητες θα αναλύσουμε το δίκαιο της Ευρωπαϊκής Ένωσης και τον τρόπο κατασκευής του δικτυακού μοντέλου αναπαράστασης αυτού.

4.2.1 Δεδομένα Δικαίου

Το δίκαιο της Ευρωπαϊκής Ένωσης (Ε.Ε.) αποτελείται από τις ιδρυτικές συνθήκες και τη νομοθεσία, όπως οι κανονισμοί και οι οδηγίες, που έχουν άμεση ή έμμεση επίδραση στις νομοθεσίες των κρατών μελών της Ε.Ε.. Υπάρχουν τρεις πηγές δικαίου στην Ε.Ε.:

1. **Πρωτογενές**, οι Συνθήκες για την ίδρυση της Ε.Ε..

2. **Δευτερογενές**, κανονισμοί, οδηγίες και άλλες νομοθετικές πράξεις, οι οποίες βασίζονται στις Συνθήκες.

3. **Συμπληρωματική Νομοθεσία**, η νομολογία του δικαστηρίου των Ευρωπαϊκών Κοινοτήτων, το διεθνές δίκαιο και οι γενικές αρχές του δικαίου.

Το δίκαιο της Ε.Ε. είναι προσβάσιμο από το κοινό μέσω του EUR-Lex[1], την επίσημη νομική πύλη της Ευρωπαϊκής Κοινότητας. Η πύλη EUR-Lex περιέχει όλα τα έγγραφα που

[1] `http://eur-lex.europa.eu`

έχουν εκτυπωθεί στην Επίσημη Εφημερίδα της Ε.Ε. από το 1951. Για τους σκοπούς της εργασίας μας, κατεβάσαμε όλα τα έγγραφα, αφαιρώντας περιττή html μορφοποίηση, για να αποκτήσουμε ένα αντίγραφο των κειμένων της νομικής βάσης δεδομένων της Ε.Ε..

Η βάση δεδομένων είναι οργανωμένη σε τομείς, όπως περιγράφεται στον Πίνακα 4.1, μαζί με τον αριθμό των εγγράφων σε κάθε τομέα, μέχρι τον Ιούλιο του 2013. Χρησιμοποιούμε το σύνολο της νομοθεσίας, Τομείς 1 έως 6 της βάσης δεδομένων, σύμφωνα με τις τρεις πηγές του δικαίου της Ε.Ε., σχηματίζοντας μια συλλογή 249.690 νομικών εγγράφων.

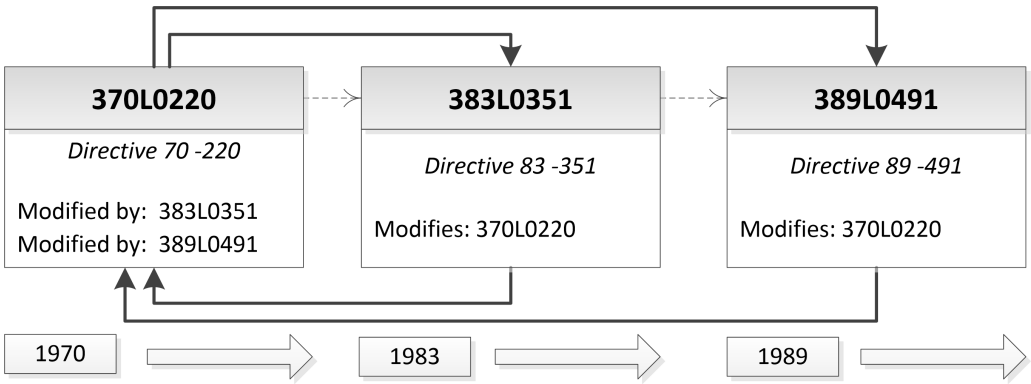
Πίνακας 4.1: Ταξινόμηση νομικών εγγράφων σε τομείς EUR-Lex. Ο αρ. εγρ. αντιστοιχεί στον αριθμό των εγγράφων σε κάθε τομέα, μέχρι τον Ιούλιο του 2013.

Ταξινόμηση EUR-Lex			
Τομέας	Τίτλος	Περιγραφή	Αρ. Εγρ.
1	Συνθήκες	Συνθήκες για την ίδρυση της Ε.Ε./ συμπληρωματικές συνθήκες	8652
2	Διεθνείς συμφωνίες	Συμφωνίες μεταξύ της Ε.Ε. και άλλων χωρών	8564
3	Παράγωγο δίκαιο	Δευτερεύουσα νομοθεσία για την εφαρμογή της πολιτικής της Ε.Ε.	120,550
4	Συμπληρωματική νομοθεσία	Συμφωνίες μεταξύ κρατών μελών	1231
5	Προπαρασκευαστικές πράξεις	Προτάσεις για μελλοντική νομοθεσία / γνώμες	73,123
6	Νομολογία	Νομολογία (αποφάσεις, διαταγές, ερμηνείες και άλλες πράξεις)	37,570
		ΣΥΝΟΛΟ	249,690

Η βάση δεδομένων EUR-Lex παρέχει αναλυτικά μεταδεδομένα για κάθε έγγραφο. Οι βιβλιογραφικές σημειώσεις ενός εγγράφου περιέχουν πληροφορίες όπως ημερομηνίες ισχύος και εγκυρότητας, η νομική μορφή του εγγράφου, συγγραφείς, το αντικείμενο, το νομικό έγγραφο από το οποίο το έγγραφο αντλεί την εξουσία του, καθώς και διάφορες σχέσεις με άλλα έγγραφα και ταξινομήσεις.

Τα πεδία, τα οποία παρέχουν συνδέσεις με άλλα έγγραφα στη βάση δεδομένων, έχουν ιδιαίτερη σημασία για τη μελέτη μας. Το Σχήμα 4.1 οπτικοποιεί ένα παράδειγμα αναπαράστασης μιας ακολουθίας τροποποιήσεων που υφίσταται ένα νομικό έγγραφο με τη μορφή τροπολογιών. Η οδηγία του Συμβουλίου 370L0220, με ημερομηνία έκδοσης 20 Μαρτίου 1970, *τροποποιήθηκε από* την οδηγία 383L0351 στις 16 Ιουνίου 1983 και στη συνέχεια *τροποποιήθηκε από* την οδηγία 389L0491 στις 17 Ιουλίου 1989. Αυτή η ακολουθία τροποποιήσεων είναι μια αμφίδρομη σχέση. Αυτό οφείλεται στο γεγονός ότι, δεδομένου ότι η οδηγία 383L0351 *τροποποιεί* την οδηγία 370L0220 και επομένως η οδηγία 370L0220 *τροποποιείται από* την οδηγία 383L0351.

Οι αναφορές στη νομοθεσία μπορούν να χωριστούν σε δύο διαφορετικές κατηγορίες: (α) μόνο για ανάγνωση αναφορές που δεν τροποποιούν το έγγραφο προορισμού και (β) αναφορές επεξεργασίας, οι οποίες τροποποιούν είτε το κείμενο ή τη διάρκεια του κύκλου ζωής του εγγράφου προορισμού. *Γενικές παραπομπές* είναι ένα παράδειγμα της πρώτης περίπτωσης, ενώ



Σχήμα 4.1: Συνδέσεις παραπομπής μεταξύ των νομικών εγγράφων. *Τροποποιήθηκε από* και *Τροποποίηση* του είναι αμφίδρομες σχέσεις.

τροποποιήθηκε με είναι ένα παράδειγμα της δεύτερης. Η εσωτερική αναφορά είναι μια αναφορά που παραπέμπει σε ένα άρθρο του ίδιου εγγράφου και, κατά συνέπεια, εξαιρείται από το πεδίο εφαρμογής της μελέτης μας.

Πίνακας 4.2: Τύποι αναφορών Νομικών Εγγράφων

Τύπος		Περιγραφή	%
Τροποποιήθηκε από	Amended by	Το έγγραφο τροποποιήθηκε από κάποιο άλλο	9,50%
Τροποποίηση του	Amendment to	Το έγγραφο τροποποιεί ένα άλλο έγγραφο	9,50%
Νομική Βάση	Legal basis	Το έγγραφο εξουσιοδοτείται από το έγγραφο	23,50%
Γενική Παραπομπή	Instruments cited	Το έγγραφο αναφέρει άλλα παρόμοια κείμενα	54,93%
Τροποποίηση από Νομολογία	Affected by case	Το έγγραφο τροποποιήθηκε ως αποτέλεσμα νομολογίας	2,00%
Λοιπά	Other	Διάφοροι τύποι αναφορών.	0,57%

Ο Πίνακας 4.2 παρέχει μια επισκόπηση της κατηγοριοποίησης των τύπων αναφορών που υπάρχουν στη βάση δεδομένων του γραπτού δικαίου της Ε.Ε.. Ο τύπος αναφοράς Γενική Παραπομπή - Instruments cited συναντάται περισσότερο από το μισό, κοντά στο 55%, επί του συνόλου των αναφορών. Αυτό σημαίνει ότι αν θεωρήσουμε την νομοθεσία ως ένα απλό δίκτυο αναφορών (citation network), όπως οι προηγούμενες μελέτες, τότε θα έχουμε αγνοήσει σχεδόν το 45% του συνόλου των σχέσεων. Είναι σαφές ότι οι προηγούμενες μελέτες που επικεντρώνονται αποκλειστικά στην ανάλυση παραπομπών επί νομικών σωμάτων, αγνοούν ένα σημαντικό ποσοστό από τις ιδιότητες του δικτύου δικαίου.

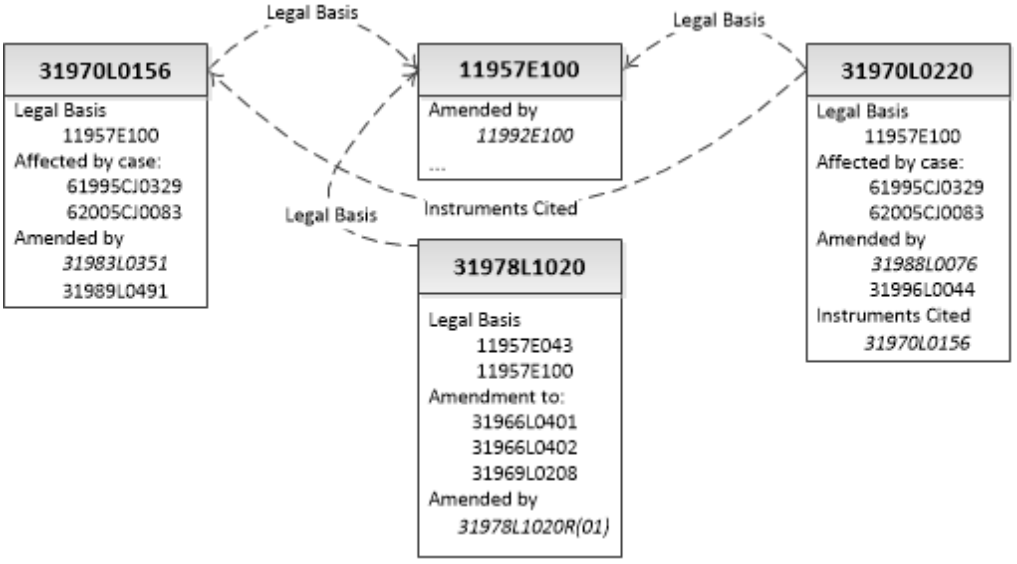
4.2.2 Κατασκευή Δικτύου Νομοθεσίας

Σε γενικές γραμμές, το γραπτό δίκαιο αποτελείται από έναν αριθμό κανονιστικών εγγράφων που παραπέμπουν σε άλλα. Με τον τρόπο αυτό μπορεί να κατασκευαστεί ένα κατευθυνόμενο

δίκτυο, καθώς ένα νομικό έγγραφο δύναται να παραπέμπει σε ένα άλλο έγγραφο (εξερχόμενη ακμή), ή να παραπέμπεται από ένα άλλο έγγραφο (εισερχόμενη ακμή). Επιπλέον, δεδομένου ότι τα νομικά έγγραφα μπορούν να αναφερθούν σε άλλα υπάρχοντα νομικά έγγραφα, το μοντέλο μας είναι στην πραγματικότητα ένα κατευθυνόμενα ακυκλικό γράφημα (DAG).

Οι τύποι των κόμβων του δικτύου ποικίλλουν ανάλογα με τον αντίστοιχους τομείς των νομικών εγγράφων, όπως προαναφέρθηκε στον Πίνακα 4.1. Οι ακμές του γραφήματος έχουν τύπους ανάλογα με το είδος των αναφορών που υπάρχουν στη βάση δεδομένων EUR-Lex, όπως απεικονίζεται στον Πίνακα 4.2. Συνολικά το δίκτυο αποτελείται από 234.287 κόμβους και 998.595 ακμές που συνδέουν τους κόμβους.

Οι κόμβοι και οι ακμές του δικτύου Νομοθεσίας έχουν χρονικά χαρακτηριστικά επίσης. Κάθε κόμβος έχει επισημανθεί με *ημερομηνία έναρξης ισχύος*, την ημερομηνία κατά την οποία η νομοθεσία τέθηκε σε ισχύ και *ημερομηνία λήξης*, την ημερομηνία κατά την οποία η νομοθεσία θα πάψει να έχει εφαρμογή. Αρκετά συχνά η νομοθεσία εγκρίνεται χωρίς ρητά οριζόμενη ημερομηνία λήξης, που ονομάζεται επίσης ως *sunset close*. Για τις νομικές πηγές χωρίς sunset close, θέσαμε ημερομηνία λήξης το έτος 9999. Οι ακμές του δικτύου ακολουθούν την χρονική κατανομή των αντίστοιχων κόμβων. Δηλαδή, μια ακμή υφίσταται μόνο για τις χρονικές περιόδους μεταξύ των ημερομηνιών έναρξης ισχύος και λήξης που προκύπτουν από τους κόμβους που συνδέουν. Αυτό το χαρακτηριστικό γνώρισμα του Δικτύου Νομοθεσία έχει ιδιαίτερη σημασία καθώς μας επιτρέπει να αναπαράγουμε την ισχύουσα νομοθεσία σε κάθε δεδομένη χρονική στιγμή.



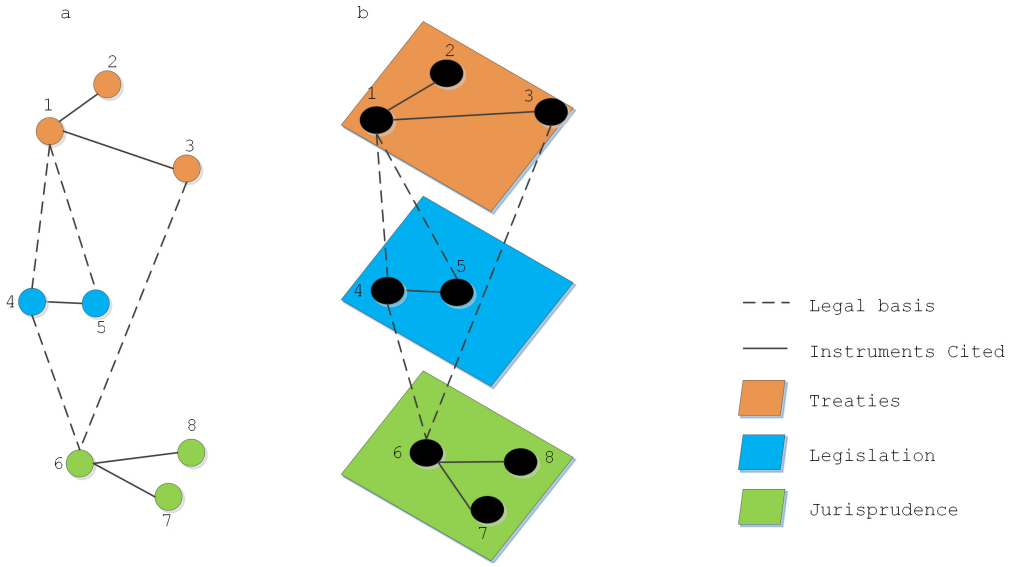
Σχήμα 4.2: Σχηματισμός του Δικτύου νομοθεσίας.

Το Σχήμα 4.2 παρουσιάζει το σχηματισμό του δικτύου του δικαίου της Ε.Ε. από τη νομική βάση δεδομένων. Κόμβοι του δικτύου είναι τα νομικά έγγραφα. Κάθε έγγραφο της νομικής συλλογής αναλύεται για παραπομπές. Εάν βρεθεί μια παραπομπή μεταξύ δύο εγγράφων, τότε η κατάλληλου τύπου ακμή συνδέει αυτούς τους δύο κόμβους.

Το Δίκτυο Νομοθεσίας της Ε.Ε., όπως πολλά δίκτυα στον πραγματικό κόσμο, παρου-

σιάζει τόσο διαχρονική εξέλιξη όσο και σύνθετη δομή πολλαπλών επιπέδων [69]. Είναι ένα πολυεπίπεδο δίκτυο, με ετερογενές σύνολο ακμών, τα οποία αντιπροσωπεύουν τις αναφορές των διαφόρων τύπων. Επίσης αποτελείται από διαφορετικούς τύπους κόμβων, ξεχωριστής σημασίας στο σώμα του γραπτού δικαίου, σύμφωνα με τις τρεις πηγές του δικαίου της Ε.Ε.. Κάθε μία από αυτές τις πηγές αποτελεί ένα υπό-δίκτυο στο οποίο οι κόμβοι συνδέονται μεταξύ τους, αλλά και σε άλλα υπό-δίκτυα.

Μια οπτική αναπαράσταση του Δικτύου Νομοθεσίας παρέχεται στο Σχήμα 4.3. Διάφορες νομικές πηγές: Συνθήκες (Treaties)(κόκκινοι κόμβοι/ πρώτο επίπεδο), Διεθνείς συμφωνίες (International agreements) (μπλε κόμβοι/ δεύτερο επίπεδο) και Νομοθεσία (Legislation)(πράσινοι κόμβοι/ τρίτο επίπεδο) που συνδέονται με ακμές του τύπου Νομική Βάση (Legal basis) (διακεκομμένη γραμμή) και Γενική Παραπομπή (Instruments cited) (συνεχής γραμμή).



Σχήμα 4.3: (α) Ένα παράδειγμα της δομής πολλαπλών επιπέδων του Δικτύου Νομοθεσίας. Τα νομικά έγγραφα που ανήκουν σε διαφορετικούς τομείς (αντιπροσωπεύονται με διαφορετικά χρώματα) διασυνδέονται με διαφορετικούς τύπους σχέσεων (Νομική Βάση, Γενική Παραπομπή). (β) Αναπαράσταση του ίδιου δικτύου, χρησιμοποιώντας επίπεδα. (Καλύτερη απεικόνιση έγχρωμα.)

Παράλληλα με το σύνολο του δικτύου, Δίκτυο Νομοθεσίας - **Legislation Network (LN)**, εντοπίσαμε τα ακόλουθα επιμέρους δίκτυα, τα οποία και θα εξετάσουμε σε αυτό το κεφάλαιο:

– Το υπό-δίκτυο Κανονισμών, **Regulations (RN)**. Στο υπό-δίκτυο αυτό περιλαμβάνονται μόνο τα νομικά έγγραφα του τομέα 3 'Παράγωγο Δίκαιο' (Πίνακας 4.1). Το υπό-δίκτυο αυτό ενσωματώνει την δευτερογενή νομοθεσία για την εφαρμογή της πολιτικής της Ε.Ε., με άμεση ή έμμεση επίδραση στα κράτη μέλη της Ε.Ε., καθώς περιέχει το σύνολο των: (α) κανονισμών της Ε.Ε., οι οποίοι είναι γενικής εφαρμογής, δεσμευτικοί στο σύνολό τους με άμεση ισχύ, (β) οδηγιών, οι οποίες είναι δεσμευτικές, ως προς

το αποτέλεσμα που πρέπει να επιτευχθεί, από κάποιο ή όλα τα κράτη μέλη στα οποία απευθύνονται και (γ) αποφάσεις που επίσης είναι δεσμευτικές στο σύνολό τους.

– Το υπό-δίκτυο Παραπομπών, **Instruments cited (ICN)**. Το υπό-δίκτυο των Παραπομπών περιέχει όλα τα έγγραφα του Δικτύου Νομοθεσίας και μόνον τις ακμές του τύπου Instruments cited (Πίνακας 4.2). Παρομοιάζει ένα δίκτυο τύπου παραπομπών citation network, όπως έχει προταθεί στην βιβλιογραφία.

– Το υπό δίκτυο Νομικής Βάσης, **Legal basis (LBN)**. Σε αυτό το υπό-δίκτυο υπάρχουν, μεταξύ των εγγράφων του Δικτύου Νομοθεσίας, μόνο συνδέσεις του τύπου Νομική Βάση Legal basis. Μια ακμή προστίθεται στο δίκτυο από τον κόμβο νομικό έγγραφο Α προς τον κόμβο νομικό έγγραφο Β αν το έγγραφο Α είναι εξουσιοδοτημένο από το Β. Το δίκτυο αυτό έχει μεγάλη σημασία για τον εντοπισμό της εσωτερικής ιεραρχίας της νομοθεσίας.

Υπό-δίκτυα του Δικτύου Νομοθεσίας μπορούν να κατασκευαστούν χρησιμοποιώντας τα ακόλουθα κριτήρια: τύπος τομέα, τύπος αναφοράς, χρονικό διάστημα ή ακόμα και συνδυασμό αυτών. Χρησιμοποιώντας τεχνικές φιλτραρίσματος κόμβων και ακμών στο Δίκτυο Νομοθεσίας μπορεί κανείς να εντοπίσει και να εξετάσει διεξοδικά διάφορα επιμέρους υπό-δίκτυα. Ο Αλγόριθμος 4.1, χωρίζει το Δίκτυο Νομοθεσίας σε υπό-δίκτυα συγκεκριμένου τομέα της νομοθεσίας. Οι τομείς των κόμβων του δικτύου παρουσιάζονται στον πίνακα 4.1.

Αλγόριθμος 4.1 Κατασκευή υπο-δικτύων Νομοθεσίας ανά τομέα.

Input: legislation graph G, legislation sector s
Output: legislation graph G of specific sector
  $sectors \leftarrow$ list of legislation sectors
  for all $sector \in sector$ **do**
    if $s \neq sector$ **then**
      $n \leftarrow$ nodes in G of sector type s
      $e \leftarrow \text{edges}(n)$
      $G \leftarrow G \subset (n, e)$
    end if
  end for
  return G

Με παρόμοιο τρόπο, ο Αλγόριθμος 4.2, διαχωρίζει το δίκτυο της νομοθεσίας σε υπό-δίκτυα συγκεκριμένων τύπων συσχετίσεων. Τα εφαρμοστέα είδη συσχετίσεων παρουσιάζονται στον πίνακα 4.2.

Η πρόσβαση στη νομοθεσία ανακτά, συνήθως, την τρέχουσα ισχύουσα νομοθεσία σχετικά με ένα θέμα. Τα νομικά συστήματα τύπου point-in-time, λειτουργούν με διαφορετική οπτική: οι δικηγόροι, οι δικαστές, αλλά και απλοί πολίτες, πολλές φορές αξιολογώντας τις νομικές συνέπειες των γεγονότων του παρελθόντος, θα πρέπει να γνωρίζουν την ισχύουσα νομοθεσία σε κάποιο σημείο του παρελθόντος, όταν συνέβησαν γεγονότα που οδήγησαν σε διένεξη ή σε δικαστικές διαμάχες [153]. Ο Αλγόριθμος 4.3 μας δίνει τη δυνατότητα να αναπαραστήσουμε το δίκτυο της ισχύουσας νομοθεσίας σε συγκεκριμένο χρονικό διάστημα.

Αλγόριθμος 4.2 Κατασκευή υπο-δικτύων Νομοθεσίας ανά τύπο νομικής παραπομπής.

Input: legislation graph G, relation type r
Output: legislation graph G of specific relation
  $relations \leftarrow$ list of legislation relations
  for all $relation \in relations$ **do**
    if $r \neq relation$ **then**
      $e \leftarrow$ edges in G of relation type r
      $G \leftarrow G \subset (e)$
    end if
  end for
  return G

Αλγόριθμος 4.3 Κατασκευή ισχύουσας Νομοθεσίας σε συγκεκριμένο χρονικό διάστημα.

Input: Complete legislation graph G, time step period t
Output: legislation (in effect) graph G for time period t
  $n1 \leftarrow$ expired nodes in G ▷ date of $expiration < t$
  $e1 \leftarrow$ edges$(n1)$
  $n2 \leftarrow$ (future) nodes in G, ▷ date of affect $> t$
  $e2 \leftarrow$ edges$(n2)$
  $G \leftarrow G \subset (n1, n2, e1, e2)$
  return G

Ο Πίνακας 4.3 συνοψίζει βασικές ιδιότητες του Δικτύου Νομοθεσίας και των επιμέρους υπό-δικτύων, τα οποία αναλύονται περαιτέρω. Για κάθε δίκτυο αναφέρεται ο αριθμός των κόμβων, ο αριθμός των ακμών, ο μέσος βαθμός, η διάμετρος, το μέσο μήκος διαδρομής, το μέγεθος της γιγάντιας συνιστώσας (g.c.) και ο αριθμός των απομονωμένων κόμβων. Επιπλέον, με τη χρήση παρενθέσεων, παραθέτουμε τις μετρήσεις για την τρέχουσα/ ισχύουσα έκδοση των έκαστων δικτύων. Δηλαδή, χρησιμοποιώντας τον αλγόριθμο 4.3 σχηματίζουμε την τρέχουσα (ισχύουσα) έκδοση του κάθε δικτύου και μετράμε τις προαναφερθείσες ιδιότητες.

Πίνακας 4.3: Βασικές ιδιότητες του Δικτύου Νομοθεσίας και των υπό-δικτύων. Τα νούμερα εντός παρενθέσεων, αφορούν την τρέχουσα / ισχύουσα έκδοση έκαστων δικτύων.

Metrics/ Network	Legislation (LN)	Regulations (RN)	Inst. cited (ICN)	Legal basis (LBN)
# of nodes	234,287 (122,091)	115,105 (36,330)	140,208 (85,417)	163,095 (51,898)
# of edges	998,595 (524,503)	338,134 (72,605)	554,917 (387,803)	237,531 (74,467)
Average degree	8.52 (8.59)	5.88 (4.00)	7.92 (9.08)	2.91 (2.87)
Network diameter	39 (33)	41 (30)	79 (60)	6 (6)
Average path length	7.22 (7.58)	7.00 (7.20)	7.54 (6.90)	1.66 (1:48)
Size of g.c.	233,337 (116,790)	112,532 (29,583)	133,211 (78,140)	161,081 (49,038)
% of g.c	99.6% (95.7%)	97.8% (81.4%)	95% (91.5%)	98.8% (94.5%)
Isolated nodes	950 (5,301)	2,573 (6,747)	6,997 (7,277)	2,014 (2,860)

Η διάμετρος και το μέσο μήκος διαδρομών του υπό-δικτύου Παραπομπών είναι αναλογικά υψηλότερα σε σύγκριση με τα άλλα δίκτυα. Αντίθετα, ο μέσος βαθμός και το μέσο μήκος διαδρομών του υπό-δικτύου Νομικής Βάσης είναι αρκετά μικρότερα σε σχέση με τα άλλα δίκτυα, καθώς το υπό-δίκτυο αυτό περιλαμβάνει μόνο τις ακμές του τύπου νομική βάση. Επιπλέον,

σημειώνουμε ότι η τρέχουσα (ενεργή) έκδοση του κάθε υπό-δικτύου είναι λιγότερο συνδεδεμένη, καθώς ένα μικρότερο ποσοστό των κόμβων ανήκουν στην γιγάντια συνιστώσα και, κατά συνέπεια, υπάρχουν περισσότεροι απομονωμένοι κόμβοι.

4.3 Ανάλυση Δικτύου Νομοθεσίας

Σε αυτή την ενότητα εφαρμόζουμε την μοντελοποίηση σύνθετου δικτύου του Δικαίου της Ε.Ε. Αρχικά θα εξετάσουμε την δομή και τα τοπολογικά χαρακτηριστικά του, προσπαθώντας να εντοπίσουμε οργανωτικές αρχές της νομοθεσίας, ιδιότητες και συμπεριφορές, οι οποίες θα παρέμεναν κρυφές εάν κάποιος μελετούσε αποσπασματικά τα νομικά έγγραφα. Στην συνέχεια, μελετάμε πώς το γραπτό δίκαιο εξελίσσεται με την πάροδο του χρόνου, καθώς καινούργια έγγραφα τίθενται σε ισχύ ενώ άλλα τροποποιούνται ή καταργούνται. Τέλος, αξιολογούμε την ανοχή του Δικτύου Νομοθεσίας σε αλλαγές, πραγματοποιώντας μια δοκιμή ανθεκτικότητας.

4.3.1 Δομή και Τοπολογικά Χαρακτηριστικά

Κρίσιμο ζήτημα για την κατανόηση της λειτουργίας ενός πολύπλοκου συστήματος, είναι ο χαρακτηρισμός των δομικών ιδιοτήτων του υποκείμενου δικτύου [11]. Ένα σημαντικό εύρημα της ανάλυσης δικτύων αποτελεί το γεγονός ότι τα δίκτυα σε φυσικά, τεχνολογικά και κοινωνικά συστήματα δεν είναι τυχαία, αλλά ακολουθούν μια σειρά από βασικές οργανωτικές αρχές ως προς τη δομή και την εξέλιξη τους, διαφοροποιώντας τα από τυχαία συνδέδεμένα δίκτυα [5]. Η ανάλυση της δομής ενός δικτύου εξετάζει τόσο από μακροσκοπική όσο και από μικροσκοπική πλευρά την τοπολογία ενός δικτύου. Οι μακροσκοπικές παρατηρήσεις περιγράφουν την γενικότερη δομή του δικτύου και μας επιτρέπουν να ερμηνεύσουμε την επίδραση της δομής του δικτύου σε μεμονωμένους κόμβους στο δίκτυο. Αντιθέτως, οι μικροσκοπικές παρατηρήσεις ποσοτικοποιούν την σχετική σημασία ενός κόμβου στο δίκτυο και μας επιτρέπουν να ερμηνεύσουμε την επιρροή μεμονωμένων κόμβων στην γενική δομή του δικτύου.

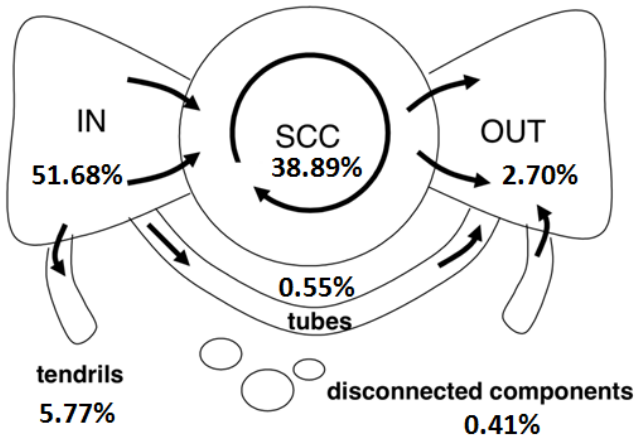
Δομή του Δικτύου

Ένα δημοφιλές μοντέλο για την οπτικοποίηση της μακροσκοπική δομής κατευθυνόμενων δικτύων είναι το μοντέλο bow-tie π.χ., η δομή bow-tie του παγκόσμιου ιστού [23]. Το Σχήμα 4.4 απεικονίζει τη δομή bow-tie του Δικτύου Νομοθεσίας, ενώ τα μεγέθη των διαφόρων στοιχείων δίνονται στον πίνακα 4.4.

Η μακροσκοπική δομή του Δικτύου Νομοθεσίας, σύμφωνα με το μοντέλο bow-tie, αποτελείται από τα ακόλουθα συστατικά:

- SCC, Ο πυρήνας, το κύριο συστατικό, είναι η ισχυρά συνδεδεμένη συνιστώσα που ονομάζεται SCC και περιέχει όλα τα νομικά έγγραφα που συνδέονται μεταξύ τους με κατευθυνόμενες ακμές.

- IN, περιέχει νομικά έγγραφα που παραπέμπουν στην ισχυρά συνδεδεμένη συνιστώσα, αλλά όχι το ανάποδο



Σχήμα 4.4: Η Δομή του Δικτύου Νομοθεσίας με το μοντέλο bow-tie. Κάθε νομικό έγγραφο του που ανήκει στο IN, συνδέεται μέσω του SCC με τα έγγραφα του OUT. Προσαρτημένα στα IN και OUT είναι τα TENDRILS που περιλαμβάνουν νομικά έγγραφα που είναι προσβάσιμα από τμήματα του IN ή έχουν πρόσβαση σε τμήματα του OUT αντίστοιχα, χωρίς να χρειαστεί να αναφερθούν σε έγγραφα του SCC. Είναι πιθανό ένα TENDRIL που βρίσκεται κάτω από το IN, να συνδεθεί με TENDRIL που βρίσκεται κάτω από το OUT, σχηματίζοντας ένα TUBE - ένα πέρασμα από ένα τμήμα του IN σε ένα τμήμα του OUT χωρίς την χρήση του SCC.

Πίνακας 4.4: Μεγέθη συστατικών μοντέλου bow-tie του Δικτύου Νομοθεσίας

Component	# of nodes	% of nodes
SCC	91,107	38.89%
IN	121,093	51.68%
OUT	6,314	2.70%
TUBES	1,300	0.55%
TENDRILS	13,523	5.77%
OTHERS	950	0.41%

– OUT, αποτελείται από νομικά έγγραφα τα οποία παραπέμπονται από τα έγγραφα του πυρήνα, αλλά δεν παραπέμπουν σε αυτόν.

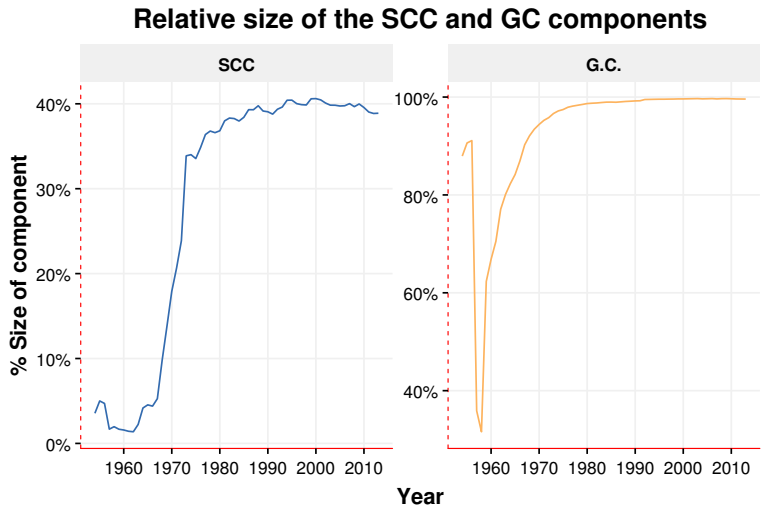
– TUBES, Τα TUBES σχηματίζονται από νομικά έγγραφα που δεν ανήκουν στον πυρήνα, είναι προσβάσιμα από το IN και έχουν πρόσβαση στο OUT.

– TENDRILS, Τα TENDRILS περιλαμβάνουν νομικά έγγραφα που είναι προσβάσιμα από τμήματα του IN ή έχουν πρόσβαση σε τμήματα του OUT αντίστοιχα, χωρίς να χρειαστεί να αναφερθούν σε έγγραφα του SCC.

– DISCONNECTED, τα υπόλοιπα νομικά έγγραφα που δεν είναι συνδεδεμένα.

Παρατηρώντας τη μακροσκοπική δομή του δικτύου νομοθεσίας σημειώνουμε ότι η ισχυρά συνδεδεμένη συνιστώσα και η συνιστώσα IN είναι σημαντικά μεγαλύτερες και η συνιστώσα OUT σημαντικά μικρότερη, σε σύγκριση με άλλες μελέτες π.χ.,του παγκόσμιου ιστού [23]. Σχεδόν όλοι οι κόμβοι ανήκουν στην γιγάντια συνιστώσα (0,41% αποσυνδεδεμένοι κόμβοι),

περίπου το 40% των εγγράφων ανήκουν στην ισχυρά συνδεδεμένη συνιστώσα και το 50% των εγγράφων μπορεί να φτάσει στον πυρήνα με άμεση διαδρομή. Γενικά το Δίκτυο Νομοθεσίας είναι καλά διασυνδεδεμένο. Τα περισσότερα νομικά έγγραφα ανήκουν στη γιγάντια συνιστώσα, και από οποιοδήποτε έγγραφο είναι δυνατόν να προσπελαστούν σχεδόν όλα τα υπόλοιπα. Η παρατήρηση αυτή οφείλεται σε ένα αφανές χαρακτηριστικό της νομοθετικής διεργασίας, που δημιουργεί συνδεδεμένες νομικές πηγές. Με τον τρόπο αυτό, το περιεχόμενο κάθε νομικής πηγής μπορεί να γίνει πλήρως κατανοητό μετά την ανάγνωση πολλών διαφορετικών εγγράφων. Επιπλέον, με βάση την εξέλιξη της μακροσκοπική δομής του Δικτύου Νομοθεσίας, όπως απεικονίζεται στο Σχήμα 4.5, παρατηρούμε ότι τόσο η ισχυρά συνδεδεμένη συνιστώσα όσο και η γιγάντια συνιστώσα αυξάνουν σε μέγεθος με την πάροδο του χρόνου.

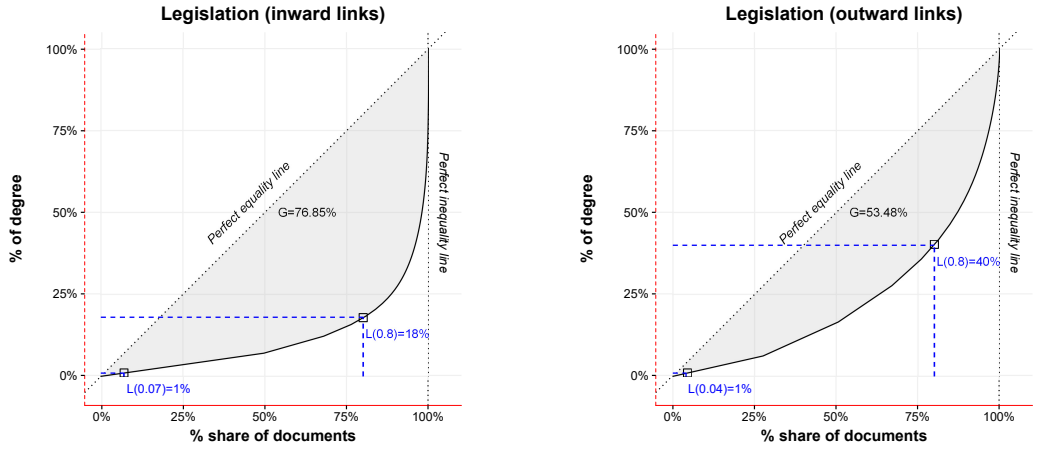


Σχήμα 4.5: Χρονική εξέλιξη μεγέθους ισχυρά συνδεδεμένης (scc) και γιγάντιας συνιστώσας (g.c)

Μια καθιερωμένη μέτρηση αναφορικά με την δομή των δικτύων είναι ο *βαθμός κατανομής*, $P(k)$, που δίνει την πιθανότητα ότι ένα τυχαία επιλεγμένο έγγραφο έχει k συνδέσμους. Ένας από τους πιο συνηθισμένους τρόπους για να περιγράψουμε την άνιση κατανομή μίας μεταβλητής είναι η καμπύλη Lorenz, η οποία χρησιμοποιήθηκε αρχικά για την περιγραφή ανισοτήτων στην οικονομική θεωρία [9]. Η καμπύλη Lorenz είναι μία γραφική αναπαράσταση της αθροιστικής συνάρτησης κατανομής μίας κατανομής πιθανότητας, ένα γράφημα που αναπαριστά το ποσοστό της κατανομής που λαμβάνεται από το κατώτερο $Y\%$ των τιμών και είναι ιδιαίτερα χρήσιμη για την γραφική αναπαράσταση δηλώσεων της μορφής 'ποσοστό $X\%$ των κόμβων συντελεί στην δημιουργία ποσοστού $Y\%$ επί των ακμών'. Για την εφαρμογή της καμπύλης Lorenz οι κόμβοι του δικτύου με τον αντίστοιχο βαθμό τους, ταξινομούνται κατά φθίνουσα σειρά του βαθμού αυτών και υπολογίζουμε την ποσοστιαία κατανομή, καθώς και την αθροιστική σειρά των ποσοστιαίων αναλογιών. Στη συνέχεια σε σύστημα ορθογώνιων συντεταγμένων απεικονίζεται η γραφική παράσταση της καμπύλης Lorenz, στην οποία ο άξονας των τετμημένων μετρά την αθροιστική σειρά της ποσοστιαίας κατανομής των κόμβων, ο δε άξονας των

τεταγμένων την αθροιστική σειρά της ποσοστιαίας κατανομής των βαθμών τους και εντοπίζονται τα σημεία της αντιστοιχίας των αθροιστικών σειρών (%) βαθμών και κόμβων, τα οποία σχηματίζουν μια πολυγωνική γραμμή, την καμπύλη Lorenz. Στην γραφική αυτή παράσταση εάν η κατανομή βαθμών των κόμβων ήταν ομοιόμορφη, η αθροιστική καμπύλη των κόμβων θα ακολουθούσε τη διαγώνιο γραμμή της αθροιστικής σειράς των βαθμών, τη γραμμή των 45°, που ονομάζεται διαγώνιος ισοκατανομής. Αντίθετα, εάν οι βαθμοί των κόμβων ακολουθούσαν την μέγιστη δυνατή ανισότητα, τότε η αθροιστική καμπύλη των κόμβων θα συνέπιπτε με τον άξονα των τετμημένων. Όσο μεγαλύτερη είναι η ένταση της ανισοκατανομής τόσο περισσότερο κάμπτεται η καμπύλη προς τα κάτω και δεξιά της διαγωνίου.

Ένας βασικός δείκτης μέτρησης ανισότητας, ο συντελεστής Gini χρησιμοποιείται κυρίως στην οικονομική θεωρία για να χαρακτηρίσει την ανισότητα που υπάρχει στην κατανομή του πλούτου, μέσω της καμπύλης Lorenz, αλλά μπορεί να χρησιμοποιηθεί για τη μέτρηση της ετερογένειας μιας εμπειρικής κατανομής. Ο συντελεστής Gini λαμβάνει τιμές μεταξύ μηδέν και ένα, με το μηδέν (0) να δηλώνει πλήρη ισότητα μεταξύ των βαθμών και το ένα (1) να δηλώνει την κυριαρχία ενός μόνο κόμβου. Όσο υψηλότερη είναι η τιμή του συντελεστή, τόσο πιο άνιση είναι η κατανομή. Ο συντελεστής Gini [57] είναι ο λόγος της περιοχής μεταξύ της διαγωνίου ισοκατανομής και της παρατηρούμενης καμπύλης Lorenz προς την περιοχή μεταξύ της διαγωνίου ισοκατανομής και της γραμμής τέλειας ανισότητας.



(α′) Δίκτυο Νομοθεσίας Εισερχόμενες Παραπομπές

(β′) Δίκτυο Νομοθεσίας Εξερχόμενες Παραπομπές

Σχήμα 4.6: Καμπύλη Lorenz και συντελεστής Gini για τον βαθμό κατανομής νομικών εγγράφων βάση εισερχομένων και εξερχομένων παραπομπών στο Δίκτυο Νομοθεσίας. Παραθέτουμε τις τιμές για το ποσοστό 1% και το κανόνα Pareto 80/20. (Καλύτερη απεικόνιση έγχρωμα.)

Το Σχήμα 4.6 παρουσιάζει την καμπύλη Lorenz μαζί με τον συντελεστή Gini για το βαθμό κατανομής κόμβων στο Δίκτυο Νομοθεσίας. Αν το Δίκτυο Νομοθεσία ήταν ένα τυχαίο Erdös-Rényi δίκτυο, η καμπύλη Lorenz θα πρέπει να είναι κοντά στην διαγώνιο ισοκατανομής και η συνάρτηση των βαθμών κατανομής θα έπρεπε να ακολουθεί μια κατανομή τύπου Poisson.

Παρατηρούμε στις γραφικές παραστάσεις ότι η καμπύλη Lorenz δεν είναι κοντά στην διαγώνιο γραμμή, αλλά αποκλίνει προς την ανισότητα. Η πλειοψηφία των εγγράφων έχουν ελάχιστες συνδέσεις, ενώ υπάρχουν και μερικά έγγραφα που είναι ευρέως συνδεδεμένα. Επίσης πολλοί κόμβοι με μικρό βαθμό συνυπάρχουν με λίγους κόμβους που έχουν εξαιρετικά μεγάλο αριθμό συνδέσεων. Το 1% των κόμβων συμμετέχουν στο 7% του συνόλου των εισερχομένων και 4% του συνόλου των εξερχομένων ακμών. Στο διάγραμμα παρουσιάζουμε επίσης τον κανόνα Pareto 80/20: ποσοστό των κόμβων με υψηλό βαθμό που αντιπροσωπεύουν το 80% όλων των συνδέσμων. Για παράδειγμα, στο Σχήμα 4.6α′ μπορούμε να δούμε ότι το 80% του συνόλου των εισερχομένων παραπομπών αποδίδεται στο 17% των νομικών εγγράφων με την υψηλότερη κατανομή βαθμών, στο Δίκτυο Νομοθεσίας, ενώ το 80% του συνόλου των εξερχομένων παραπομπών αποδίδεται στο 40% νομικών εγγράφων με την υψηλότερη κατανομή βαθμών, όπως φαίνεται στο Σχήμα 4.6β′.

Κατανομή Νόμου-Δύναμης

Σε πολλές περιπτώσεις το κλάσμα των κόμβων με βαθμό k, $P(k)$, μειώνεται ακολουθώντας μια κατανομή βασισμένη στο νόμο-δύναμης (power-law):

$$P(k) \propto K^{-\gamma} \tag{4.1}$$

όπου το γ είναι μία πραγματική σταθερά, παράμετρος της κατανομής γνωστή ως εκθέτης ή παράμετρος μείωσης, που τυπικά κυμαίνεται στα όρια $2 < \gamma < 3$. Το χαρακτηριστικό αυτό έχει παρατηρηθεί σε δίκτυα επικοινωνίων μεγάλης κλίμακας, βιολογικά και κοινωνικά συστήματα [5, 12, 103] και στο δίκτυο παραπομπών των αποφάσεων του Ανώτατου δικαστήριου των Η.Π.Α. [48, 135].

Στην πράξη, ελάχιστα εμπειρικά φαινόμενα / εμπειρικές κατανομές ακολουθούν το νόμο-δύναμης για όλες τις τιμές των x. Τις περισσότερες φορές, ο νόμος-δύναμης ισχύει μόνο για τιμές μεγαλύτερες από κάποιο ελάχιστο x_{min}. Σε τέτοιες περιπτώσεις, μπορούμε να πούμε ότι η ουρά της κατανομής ακολουθεί νόμο-δύναμης.

Ο ισχυρισμός ότι τα δεδομένα ακολουθούν πράγματι μια σχέση κατανομής νόμου δύναμης απαιτεί περισσότερα από την απλή αντιστοίχιση ενός συγκεκριμένου μοντέλου στα δεδομένα. Προκειμένου να χαρακτηρίσουμε την συνάρτηση πιθανότητας κατανομής βαθμών στο Δίκτυο Νομοθεσία, χρησιμοποιήσαμε τη μεθοδολογία που περιγράφεται στο [31]. Οι εναλλακτικές μέθοδοι βασίζονται συνήθως στη δημιουργία γραμμικής παλινδρόμησης είτε στους λογαρίθμους των διαγραμμάτων συχνοτήτων είτε σε λογαριθμημένα δεδομένα, αλλά αυτές οι προσεγγίσεις είναι καλό να αποφεύγονται καθώς μπορούν να οδηγήσουν σε ιδιαίτερα μεροληπτικές εκτιμήσεις του εκθέτη.

Συγκεκριμένα, κάθε διακριτή τιμή του βαθμού κατανομής x, αντιπροσωπεύει μια οριακή τιμή υποψήφιο, x_{min}, πάνω από την οποία η συμπεριφορά της συνάρτησης πιθανότητας, που προκύπτει με το μοντέλο του νόμου δύναμης μπορεί να παρέχει μια εύλογη προσαρμογή, για την συνάρτηση πιθανότητας κατανομής βαθμών στο Δίκτυο Νομοθεσίας. Εν συνεχεία προσδιορίζουμε, με την μέθοδο μέγιστης πιθανοφάνειας, τις παραμέτρους του μοντέλου που ελαχι-

στοποιούν το τυπικό σφάλμα της εκτίμησης για όλες τις υποψήφιες τιμές x_{min} και επιλέγουμε την ποιο ταιριαστή. Στην συνέχεια, ο έλεγχος στατιστικής υπόθεσης, για το πόσο καλά η εξίσωση που σχηματίσαμε περιγράφει τα δεδομένα μας περιλαμβάνει την σύνθεση $m = 2500$ δειγμάτων βαθμών κατανομής από το θεωρητικό μοντέλο με τιμές κατωφλίου και παραμέτρων ίσες με εκείνες που προσεγγίζονται για την υφιστάμενη κατανομή βαθμών. Για να μπορέσουμε να μετρήσουμε την τιμή στάθμη σημαντικότητας p με ακρίβεια 2 δεκαδικών ψηφίων, επιλέξαμε να δημιουργήσουμε $m = 2500$ συνθετικά σύνολα δεδομένων, όπως προτείνεται στο [31]. Το μοντέλο του νόμου δύναμης που επιλέξαμε μπορεί να αποκλειστεί, αν το 10%, $p < 0.10$, από τα συνθετικά σύνολα δεδομένων παρουσιάζει στατιστικά σημαντική διαφορά με τα πραγματικά δεδομένα.

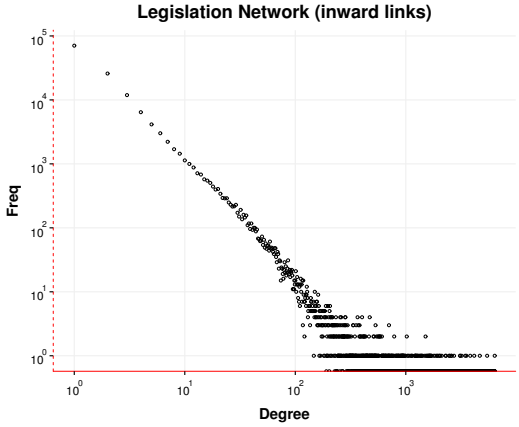
Στον πίνακα 4.5 παρουσιάζουμε τα αποτελέσματα από την εκτίμηση των παραμέτρων του νόμου δύναμης στο Δίκτυο Νομοθεσία, για εισερχόμενες και εξερχόμενες παραπομπές, καθώς και βασικές μετρήσεις της κατανομής βαθμών, όπως μέση τιμή, απόκλιση και μέγιστη τιμή. Η τελευταία στήλη του πίνακα αναφέρει την στάθμη σημαντικότητας ($p - value$) για το μοντέλο νόμου δύναμης. Για τους υπολογισμούς χρησιμοποιήσαμε το προγραμματιστικό πακέτο R poweRlaw [56].

Πίνακας 4.5: Βασικές Μετρήσεις για την συνάρτηση κατανομής βαθμών στο Δίκτυο Νομοθεσίας, όπως μέση τιμή, απόκλιση και μέγιστη τιμή, καθώς και παράμετροι του μοντέλου νόμου δύναμης. Το n υποδηλώνει το μέγεθος των κόμβων, $n, tail$ είναι το μέγεθος της ουράς του νόμου δύναμης και γ είναι ο εκθέτης της κατανομής. Οι τιμές $p - value$ προέκυψαν από bootstrap με 2500 σύνολα συνθετικών δεδομένων. (Στατιστικά σημαντικές τιμές παρουσιάζονται με **έντονη γραφή**.)

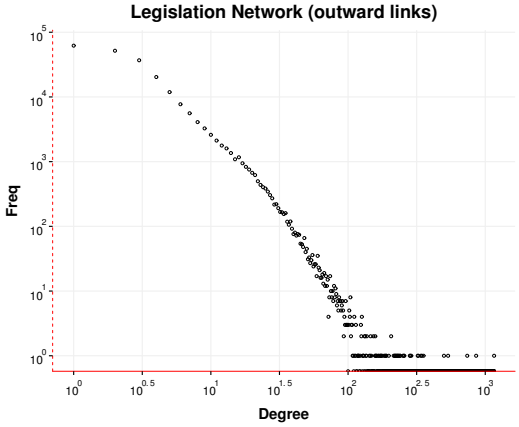
Network (Direction)	n	$\langle x \rangle$	σ	x_{max}	x_{min}	γ	n_{tail}	p
Legislation Network (In)	141,798	7.04	54.91	6373	34 ± 15	2.24 ± 0.15	4351 ± 29,700	**0.232**
Legislation Network (Out)	224,856	4.44	8.2	1160	34 ± 11	3.55 ± 0.42	2210 ± 36,000	**0.246**

Τα αποτελέσματα της ανάλυσης νόμου δύναμης απεικονίζονται στην Εικόνα 4.7. Η αριστερή στήλη αντιπροσωπεύει το υπό-δίκτυο εισερχόμενων παραπομπών, ενώ η δεξιά στήλη αντιπροσωπεύει το υπό-δίκτυο εξερχομένων. Για κάθε υπό-δίκτυο, απεικονίζουμε, σε λογαριθμική-λογαριθμική κλίμακα, στην πρώτη σειρά την κατανομή συχνοτήτων και στην δεύτερη την αθροιστική κατανομή συχνοτήτων μαζί με την κατανομή νόμου δύναμης. Σημειώνουμε ότι η απόκλιση σ είναι σημαντικά μεγαλύτερη από τις τιμές $\langle x \rangle$, αποκαλύπτοντας έτσι μεγάλες διακυμάνσεις του βαθμού κατανομής. Η πλειοψηφία των εγγράφων αναφέρονται λίγες μόνο φορές, ενώ υπάρχουν και μερικά έγγραφα που είναι ευρέως συνδεδεμένα. Η κατανομή βασισμένη στο νόμο δύναμης, με τιμές εκθέτη εισερχόμενων παραπομπών $\gamma_{in} = 2.24$ και εξερχομένων $\gamma_{out} = 3.55$, αποτελεί στατιστικά αποδεκτή κατανομή για τους βαθμούς κόμβων του Δικτύου Νομοθεσία, με τιμές $p - values$, 0.232 και 0.246 αντίστοιχα.

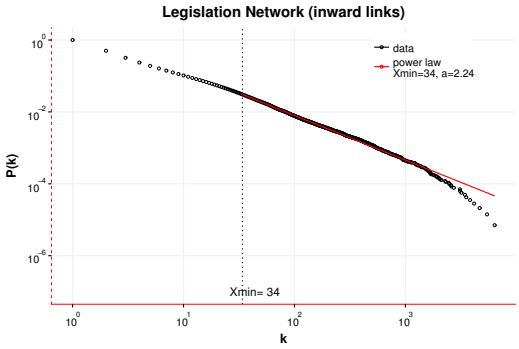
Η διαδικασία *Επιλεκτικής Προσκόλλησης* (preferential attachment), γνωστή και ως σωρευτικό πλεονέκτημα (cumulative advantage) ή 'οι πλούσιοι γίνονται πλουσιότεροι' ('the rich get richer'), που είχε προταθεί αρχικά στο πλαίσιο των διανομών πλούτου (wealth distributions) [133] και στην συνέχεια χρησιμοποιήθηκε σε πολύπλοκα δίκτυα [13], μας βοηθά να



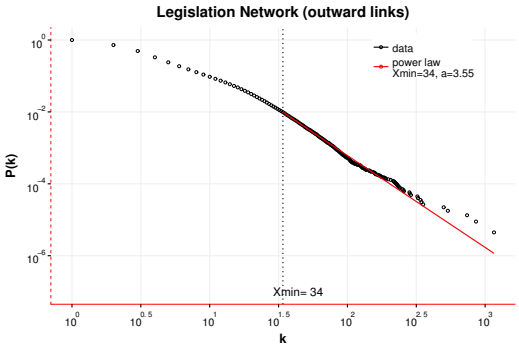
(α′) Κατανομή Συχνοτήτων Εισερχόμενες Παραπομπές



(β′) Κατανομή Συχνοτήτων Εξερχόμενες Παραπομπές



(γ′) Αθροιστική Συνάρτηση Κατανομής CDF Εισερχόμενες Παραπομπές



(δ′) Αθροιστική Συνάρτηση Κατανομής CDF Εξερχόμενες Παραπομπές

Σχήμα 4.7: Κατανομή Συχνοτήτων και Αθροιστική Κατανομή. (Καλύτερη απεικόνιση έγχρωμα.)

κατανοήσουμε την προέλευση αυτής της ετερογένειας. Κύρια θεώρηση της διαδικασίας επιλεκτικής προσκόλλησης είναι ότι η πιθανότητα σύνδεσης ενός νέου κόμβου σε υφιστάμενο είναι ανάλογη της δημοφιλίας, δηλαδή του αριθμού των συνδέσεων, του κόμβου αυτού. Επομένως οι νέοι κόμβοι προτιμούν να συνδεθούν σε ήδη καλά συνδεδεμένους κόμβους. Ειδικότερα, η ανομοιόμορφη κατανομή πιθανοτήτων προσέλκυσης νέων κόμβων, δημιουργεί δύο ομογενείς κατηγορίες κόμβων: μικρός αριθμός κόμβων με υψηλό βαθμό που λειτουργούν ως κεντρικοί κόμβοι (hubs) και η πλειοψηφία των κόμβων του δικτύου που έχουν μικρό βαθμό, οδηγώντας έτσι σε δίκτυα με ασύμμετρες κατανομές βαθμών. Επίσης στο μοντέλο αυτό, η τελική κατανομή των συνδέσεων είναι ιδιαίτερα είναι ευαίσθητη στις αρχικές συνθήκες έναρξης.

Στην μικροσκοπική θεώρηση του νομικού τομέα, όταν οι δικαστές καταγράφουν τις απόψεις τους, αναφέρουν περιπτώσεις και άλλες νομικές πηγές που θεωρούν ότι είναι σημαντικές για την περίπτωση που αποφασίζουν. Με την υποβοηθούμενη από συστήματα ανάκτηση

νομικής πληροφορίας και την ευρεία πρόσβαση σε (εμπορικές) μηχανές αναζήτησης νομικής πληροφορίας, είναι πιθανότερο να χρησιμοποιήσουν τις υψηλότερες σε κατάταξη νομικές πηγές. Οι δικηγόροι βασίζονται επίσης στις παραπομπές νομικών εγγράφων ώστε να σχηματίσουν μια καλά θεμελιωμένη νομική ανάλυση. Ως εκ τούτου, οι δικαστές και οι δικηγόροι είναι πιο πιθανό να χρησιμοποιήσουν και να παραπέμψουν, αυξάνοντας έτσι το βαθμό τους, σε έγγραφα που έχουν ήδη υψηλό βαθμό, είναι ήδη δημοφιλή.

Ομοίως, οι νομικές πηγές που βρίσκονται στην κορυφή της κατάταξης κεντρικότητας, κατατάσσονται στις πρώτες θέσεις εξ΄ αιτίας του υψηλού αριθμού συνδέσεων. Εάν ένα νομικό έγγραφο είναι ήδη δημοφιλές, είναι πιθανότερο να αποκτήσει μια καινούργια αναφορά από ένα άλλο νομικό έγγραφο. Με άλλα λόγια, διαπιστώνουμε την ύπαρξη ενός φαινομένου τύπου 'οι πλούσιοι γίνονται πλουσιότεροι' που ενισχύει τη δημοτικότητα των εγγράφων που έχουν ήδη υψηλή κατάταξη.

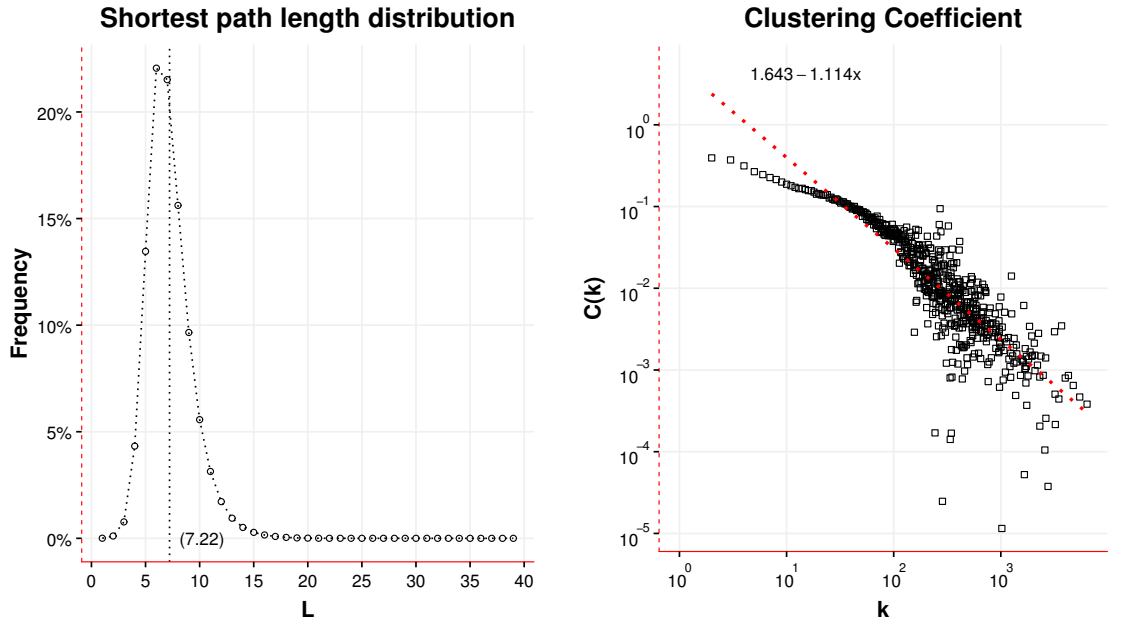
Φαινόμενο μικρού-κόσμου

Ένα άλλο σημαντικό τοπολογικό χαρακτηριστικό που παρουσιάζουν πολλά πραγματικά δίκτυα, είναι το φαινόμενο του *μικρού-κόσμου* [98]. Σύμφωνα με το [150] δίκτυα μικρού-κόσμου ορίζονται τα δίκτυα που έχουν μικρή διάμετρο και υψηλή τιμή του συντελεστή ομαδοποίησης. Πολλά κοινωνικά, τεχνολογικά, βιολογικά και πληροφορίακά δίκτυα έχουν μελετηθεί και κατηγοριοποιηθεί ως δίκτυα μικρού-κόσμου [103]. Τα δίκτυα μικρού-κόσμου μπορούν να θεωρηθούν ως συστήματα που είναι ιδιαιτέρως αποδοτικά, τόσο σε καθολικό, όσο και σε τοπικό επίπεδο, όσον αφορά την δυναμική των διεργασιών που συντελούνται σε αυτά π.χ., του πόσο αποτελεσματικά ανταλλάσσονται πληροφορίες μέσω του δικτύου [83].

Οι ιδιότητες μικρού-κόσμου μετρώνται (α) από το μέσο όρο συντομότερης διαδρομής (Μέσος αριθμός των βημάτων κατά μήκος των συντομότερων μονοπατιών για όλα τα δυνατά ζεύγη των κόμβων του δικτύου.) και (β) και το συντελεστή ομαδοποίησης, ο οποίος αποτυπώνει το βαθμό στον οποίο οι κόμβοι σε ένα δίκτυο τείνουν να συγκεντρωθούν μαζί.

Η κατανομή των συντομότερη μηκών διαδρομής απεικονίζεται στο Σχήμα 4.8α΄. Η διάμετρος D του Δικτύου Νομοθεσία, οριζόμενη ως η μεγίστη των συντομότερων μηκών διαδρομής, είναι 39 και η μέση συντομότερη διαδρομή, η μέση τιμή της γεωδαιτικής απόστασης μεταξύ ζευγαριών που έχουν τουλάχιστον ένα μονοπάτι τα που συνδέει τους, είναι $7,22$. Παρατηρούμε ότι παρά το τεράστιο μέγεθος του, το Δίκτυο Νομοθεσίας παρουσιάζει το φαινόμενο 'έξι βαθμών χωρισμού΄ (six degrees of separation).

Η κατανομή του βαθμού συντελεστής ομαδοποίησης $C(k)$, ανά βαθμό κόμβων, παρουσιάζεται στο Σχήμα 4.8β΄ σε λογαριθμική-λογαριθμική κλίμακα. Για λόγους σαφήνειας, προσθέσαμε τη γραμμή με κλίση $-1,1$. Παρά το γεγονός ότι η παρουσία νόμου δύναμης δεν μπορεί να είναι μια εύλογη προσαρμογή, ο συντελεστής ομαδοποίησης είναι αντιστρόφως ανάλογος με το βαθμό k. Οι κόμβοι υψηλού βαθμού συνδέονται με πολλούς κόμβους, που πιθανώς ανήκουν σε διαφορετικές ομάδες, με αποτέλεσμα την μικρή τιμή του συντελεστή ομαδοποίησης για τους κόμβους υψηλού βαθμού. Αντίθετα, οι κόμβοι χαμηλού βαθμού γενικά ανήκουν σε καλά διασυνδεδεμένες κοινότητες, που αντιστοιχούν σε υψηλές τιμές συντελεστή ομαδοποίησης των



(α′) Κατανομή μέσου μήκους μονοπατιών (β′) Συντελεστής ομαδοποίησης ανά βαθμό κόμβων

Σχήμα 4.8: (α) Κατανομή μέσου μήκους μονοπατιών, (β)Κατανομή συντελεστή ομαδοποίησης, $C(k)$, ανά βαθμό κόμβων, k, σε λογαριθμική-λογαριθμική κλίμακα. Από οποιοδήποτε νομικό έγγραφο είναι δυνατό να προσεγγίσουμε σχεδόν κάθε άλλο, μόλις σε 7 βήματα, κατά μέσο όρο. Ο συντελεστής ομαδοποίησης είναι αντιστρόφως ανάλογος με το βαθμό κόμβου. (Καλύτερη απεικόνιση έγχρωμα.)

κόμβων χαμηλής συνδεσιμότητας. Αυτό το μοτίβο όπως έχει μελετηθεί στο [117], υποδεικνύει την ύπαρξη μιας ιεραρχικής αρχιτεκτονικής του δικτύου. Μια ιεραρχική αρχιτεκτονική συνεπάγεται ότι οι αραιά συνδεδεμένοι κόμβοι ανήκουν σε εξαιρετικά συγκεντρωμένες περιοχές, με την επικοινωνία μεταξύ των διαφόρων ιδιαίτερα συγκεντρωμένων περιοχών να συντηρείται από λίγους κόμβους.

Για να χαρακτηριστεί ένα δίκτυο ως δίκτυο μικρού-κόσμου, συγκρίνουμε τις μετρήσεις του σε σχέση με Erdös-Rényi τυχαία δίκτυα [38], με τον ίδιο αριθμό των κόμβων και ακμών. Εάν ένα δίκτυο παρουσιάζει το φαινόμενο μικρού-κόσμου, τότε αναμένεται ότι η μέση συντομότερη διαδρομή είναι ελαφρώς μικρότερη από ενός τυχαίου δικτύου και ο μέσος συντελεστής ομαδοποίησης είναι μεγέθους μεγαλύτερος από αυτόν ενός τυχαίου δικτύου.

Παρόμοια με την βιβλιογραφία για τα δίκτυα μικρού-κόσμου [103], η ανάλυση μας περιορίζεται στην μελέτη της γιγαντιαίας συνιστώσας των δικτύων, δηλαδή το μέγιστα συνδεδεμένο υπό-γράφημα του δικτύου. Ο Πίνακας 4.6 συνοψίζει τα αποτελέσματα της ανάλυσής μας στο Δίκτυο Νομοθεσία και τις τρέχουσες (ισχύουσες) εκδόσεις της νομοθεσίας των υπό-δικτύων που εξετάζουμε. Η μέση συντομότερη διαδρομή και ο μέσος όρος των μετρήσεων συντελεστή ομαδοποίησης υποδηλώνονται από L_{net} και C_{net} και οι αντίστοιχες τιμές για τα τυχαία δίκτυα συμβολίζονται ως L_{rand} και C_{rand} αντίστοιχα.

Πίνακας 4.6: Φαινόμενο μικρού-κόσμου - Μετρήσεις μέσου όρου συντομότερης διαδρομής και συντελεστή ομαδοποίησης

	L_{net}	C_{net}	L_{rand}	C_{rand}
Legislation Network (LN)	7,22	1.1×10^{-2}	8,64	$3{,}73 \times 10^{-5}$
(current) Legislation (LN)	7,58	2.15×10^{-2}	7,97	$7{,}86 \times 10^{-5}$
(current) Regulations (RN)	7,2	7.07×10^{-3}	12,4	1.56×10^{-4}
(current) Inst. cited (ICN)	6,9	3.43×10^{-2}	7,26	1.26×10^{-4}
(current) Legal basis (LBN)	1,48	2.78×10^{-4}	24,1	$5{,}68 \times 10^{-5}$

Παρατηρούμε ότι, παρά τις διακυμάνσεις στις μετρήσεις, όλα τα δίκτυα πληρούν τις προϋποθέσεις μικρού-κόσμου. Ενδιαφέρον είναι το γεγονός ότι όχι μόνο το Δίκτυο Νομοθεσία παρουσιάζει χαρακτηριστικά μικρού-κόσμου, αλλά και οι τρέχουσες (ισχύουσες) εκδόσεις των υπο-δικτύων επίσης. Συγκρίνοντας τα αποτελέσματά μας με άλλες μελέτες, όπως αυτές παρουσιάζονται στο [103], βλέπουμε ότι οι μέσες τιμές μήκους συντομότερης διαδρομής στα επιμέρους υπό-δίκτυα νομοθεσίας είναι εμφανώς μικρότερες από τις τιμές των δικτύων που έχουν μελετηθεί στην βιβλιογραφία, αλλά και μεγέθους μικρότερες από το θεωρητικά μέσο βαθμό του αντίστοιχου τυχαίου μοντέλου. Αποδίδουμε το εύρημα αυτό στη φύση του νόμου και την ιεραρχική μορφή δομή του. Οι νομικές πηγές αντλούν την ισχύ τους από άλλες νομικές πηγές, γεγονός που μειώνει το συνολικό αριθμό των παραπομπών πολύ κάτω από τον αναμενόμενο αριθμό ενός τυχαίου μοντέλου.

Τέλος, η τάση των νομικών πηγών να αλληλεπιδρούν μεταξύ τους με συγκεκριμένο τρόπο, όπως το να συνδεθούν με άλλες πηγές που παρομοιάζουν όμοια (ή αντίθετα) χαρακτηριστικά, ονομάζεται καταταξιμότητα (assortativity) και έχει μελετηθεί στο [104]. Το επίπεδο ανάμειξης καταταξιμότητας σε ένα δίκτυο μετριέται από τον συντελεστή καταταξιμότητας, με θετικές τιμές να υποδηλώνουν ότι το δίκτυο είναι assortative, αρνητικές τιμές ότι είναι dis-assortative και μηδενικές ότι παρουσιάζεται συσχέτιση. Η καταταξιμότητα συνήθως εξετάζεται από την σκοπιά των βαθμών των κόμβων, χωρίς όμως να αποκλείεται η χρήση διαφορετικών κριτηρίων/μετρικών με βάση τα χαρακτηριστικά των υπό εξέταση δικτύων. Με βάση την κατανομή βαθμών των κόμβων τα κοινωνικά δίκτυα, παρουσιάζουν assortative ανάμειξη αφού οι κόμβοι έχουν την τάση να συνδέονται με άλλους κόμβους με παρόμοια κατανομή βαθμών, ενώ αντίθετα τα τεχνολογικά και βιολογικά δίκτυα παρουσιάζουν dissortativity, καθώς οι κόμβοι υψηλού βαθμού τείνουν να συνδέονται με κόμβους χαμηλού βαθμού.

Το Δίκτυο Νομοθεσία παρουσιάζει ένα μικρό βαθμό dis-assortative $(-0,0904)$, με κριτήριο την κατανομή βαθμών κόμβων, γεγονός που σημαίνει ότι οι κόμβοι υψηλού βαθμού (hubs) είναι πιο πιθανό να συνδεθούν με κόμβους χαμηλότερου βαθμού. Αντίθετα εξετάζοντας, τον συντελεστή καταταξιμότητας, με κριτήριο την κατανομή βαθμών κόμβων ανά τύπο νομικής πηγής, πίνακας 4.1, βλέπουμε ότι έχει τιμή 0.443, γεγονός που αποκαλύπτει ότι υπάρχει μια ισχυρή τάση των νομικών πηγών να συνδέονται με νομικές πηγές που ανήκουν στον ίδιο τύπο και να σχηματίζουν με αυτόν τον τρόπο συστάδες του ίδιου τομέα.

4.3.2 Χρονική Εξέλιξη

Τα δίκτυα που μοντελοποιούν συστήματα του πραγματικού κόσμου, εξελίσσονται με την πάροδο του χρόνου με την προσθήκη, τροποποίηση και τη διαγραφή κόμβων και ακμών. Το δίκτυο νομοθεσίας επίσης, εξελίσσεται με την πάροδο του χρόνου, με κόμβους και ακμές που προστίθενται ή απαλείφονται, καθώς καινούργια νομικά έγγραφα συνεχώς δημιουργούνται και άλλα τροποποιούνται ή ακυρώνονται. Ένα συμπληρωματικό θέμα, που συχνά αγνοείται στη βιβλιογραφία, είναι οι χρονικές πτυχές των εν λόγω δικτύων. Όπως και με την τοπολογία του δικτύου, η χρονική διάρθρωση του κόμβων και ακμών του δικτύου μπορεί να επηρεάσουν τη δυναμική των συστημάτων που αναπαραστόνται μέσω του δικτύου [123].

Στην ενότητα αυτή παρουσιάζουμε τα κύρια συμπεράσματά μας από την μελέτη της εξέλιξης της νομοθεσίας με την πάροδο του χρόνου. Προκειμένου να αξιολογηθεί η χρονική εξέλιξη του δικτύου νομοθεσία, κατασκευάζουμε υπό-δίκτυα της νομοθεσίας, σε συνεχόμενα χρονικά παράθυρα, ετησίως χωρισμένα χρονικά διαστήματα. Για κάθε χρονικό πλαίσιο ανακατασκευάζουμε το δίκτυο νομοθεσίας, με βαθμιαία προσθήκη και αφαίρεση κόμβων και ακμών, νομικών εγγράφων και αναφορών, σύμφωνα με τις ημερομηνίες έναρξης και λήξης ισχύος μέσα στο τρέχον χρονικό πλαίσιο. Για κάθε χρονική στιγμή, από το 1951 έως το 2013, δημιουργούμε υπό-δίκτυα, αναπαριστώντας την νομοθεσία που ήταν σε ισχύ ανά έκαστο χρόνο, κάνοντας χρήση του Αλγόριθμου 4.3 για την κατασκευή υπό-δικτύων ισχύουσας Νομοθεσίας σε συγκεκριμένο χρονικό διάστημα, όσο και των Αλγόριθμων 4.1 και 4.2 για την κατασκευή υπό-δικτύων Νομοθεσίας ανά τομέα και τύπο νομικής παραπομπής αντίστοιχα.

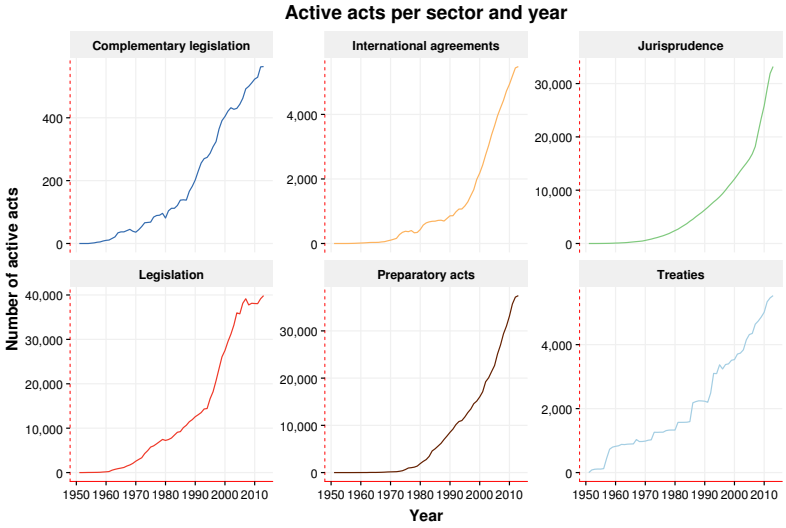
Το Σχήμα 4.9α΄ απεικονίζει την εξέλιξη του αριθμού νομικών εγγράφων σε ισχύ ανά έτος και τομέα. Ο άξονας x αντιπροσωπεύει τον χρόνο, ενώ ο άξονας y αντιπροσωπεύει τον αριθμό νομικών εγγράφων σε ισχύ. Η νομοθεσία σε ισχύ, αυξάνει με την πάροδο των ετών, σε σχέση με όλες τις πηγές δικαίου. Ομοίως, στο Σχήμα 4.9β΄, ο άξονας y αντιπροσωπεύει τον αριθμό νομικών παραπομπών κατά μήκος των διαφόρων τύπων αναφοράς. Ο αριθμός των ενεργών ακμών, των ενεργών συνδέσεων μεταξύ ενεργών νομικών εγγράφων, αυξάνει με την πάροδο των ετών σε σχέση με όλους τους τύπους νομικών παραπομπών.

Επιπρόσθετα, εξετάσαμε την εξέλιξη αυτών των μεγεθών με την πάροδο του χρόνου. Οι Leskovek et al. [85] μελέτησαν μια σειρά από διαφορετικά δίκτυα, από διάφορες περιοχές, με έμφαση στον τρόπο με τον οποίο οι θεμελιώδεις ιδιότητες του δικτύου μεταβάλλονται με το χρόνο. Κατέληξαν στο συμπέρασμα ότι ο νόμος δύναμης για την πύκνωση (densification power law) απαντάται σε ένα ευρύ φάσμα διαφορετικών δικτύων. Σύμφωνα με το νόμο δύναμης για την πύκνωση ο αριθμός των ακμών αυξάνεται ταχύτερα από τον αριθμό των κόμβων. Ειδικότερα, ο νόμος δύναμης για την πύκνωση ορίζεται από την ακόλουθη μορφή:

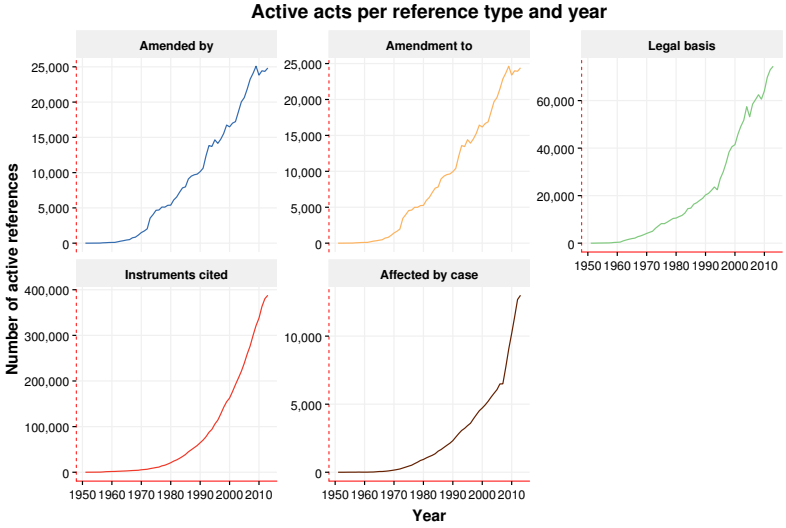
$$E(t) \propto N(t)^{\alpha} \tag{4.2}$$

όπου $E(t)$ και $N(t)$ δηλώνουν τον αριθμό των ακμών και των κόμβων του δικτύου τη χρονική στιγμή t, ενώ η πραγματική σταθερά α κυμαίνεται μεταξύ 1 και 2.

Αναλυτικότερα, η πειραματική ανάλυση μας, διεξήχθη στα τέσσερα αντιπροσωπευτικά υποδίκτυα που παρουσιάστηκαν στην Ενότητα 4.2.2. Για κάθε χρονική στιγμή t, σε ετήσια βάση,



(α΄) Αριθμός νομικών εγγράφων σε ισχύ ανά έτος και τομέα.



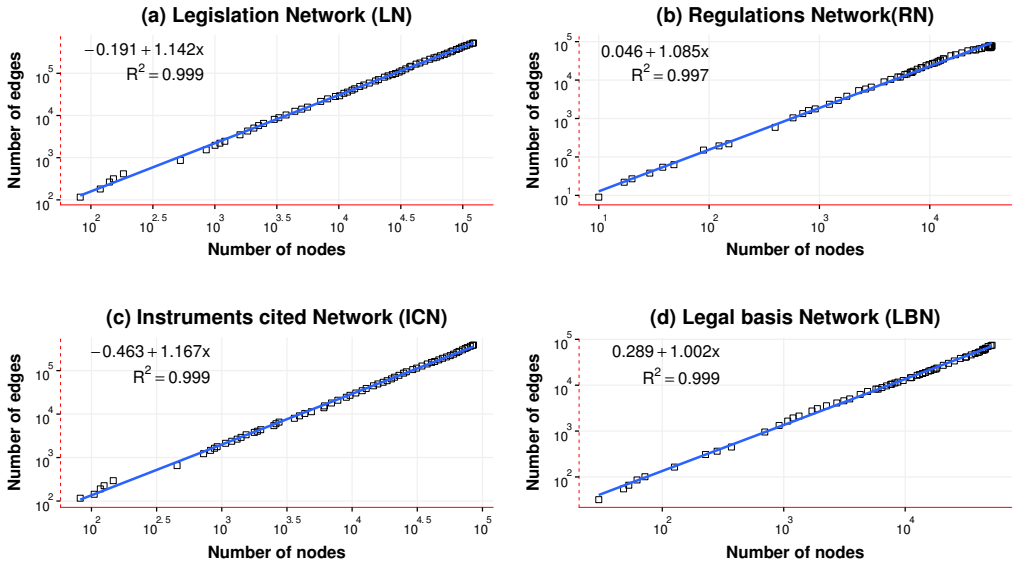
(β΄) Αριθμός νομικών παραπομπών σε ισχύ ανά έτος και ανά τύπο αναφοράς.

Σχήμα 4.9: (α) Αριθμός νομικών εγγράφων σε ισχύ ανά έτος και τομέα, (β) Αριθμός νομικών παραπομπών σε ισχύ ανά έτος και ανά τύπο αναφοράς. Ο αριθμός των νομικών εγγράφων και νομικών παραπομπών, σε ισχύ, αυξάνεται σταθερά σε όλους τους τομείς και τύπους αναφοράς. (Καλύτερη απεικόνιση έγχρωμα.)

η εκάστοτε τρέχουσα έκδοση του κάθε υπό-δίκτυου δημιουργήθηκε και αναλύθηκε. Σε συμφωνία με τα ευρήματα του [85], το Δίκτυο Νομοθεσίας ακολουθεί επίσης το νόμο δύναμης για την πύκνωση, με τον αριθμό των ακμών, των νομικών παραπομπών μεταξύ των νομικών εγγράφων, να αυξάνεται ταχύτερα από τον αριθμό των κόμβων, των νομικών εγγράφων. Σημειώνουμε ότι τα ευρήματά μας, η 'πύκνωση' του δικτύου, βασίζονται σε εξέταση του ισχύοντος Δικτύου Νομοθεσίας σε κάθε χρονικό παράθυρο και όχι σε ολόκληρη την 'στατική' εικόνα του Δικτύου Νομοθεσίας. Στην τελευταία περίπτωση αυτό είναι ένα προφανές συμπέρασμα

αφού τα νομικά έγγραφα παραπέμπουν σε υφιστάμενα νομικά έγγραφα.

Τα αποτελέσματα της χρονικής ανάλυσής παρουσιάζονται στο Σχήμα 4.10. Για κάθε υπό-δίκτυο, απεικονίζονται σε λογαριθμική-λογαριθμική (log-log) κλίμακα ο αριθμός των ενεργών ακμών και κόμβων. Ο αριθμός των εν ισχύ παραπομπών μεταξύ των νομικών εγγράφων αυξάνεται ταχύτερα από τον αριθμό των εν ισχύ νομικών εγγράφων για όλα τα υπό-δίκτυα που εξετάστηκαν. Σκοπεύουμε, σε μελλοντική εργασία, να αξιοποιήσουμε αυτό το εύρημα προς την κατεύθυνση περιγραφής ενός εξελικτικού μοντέλου, προκειμένου να μοντελοποιήσουμε εξελικτικά την διαδικασία παραγωγής νομικών εγγράφων.



Σχήμα 4.10: Γραφική αναπαράσταση του νόμου δύναμης για την πύκνωση σε υπό-δίκτυα του Δικτύου Νομοθεσίας. Αναπαράσταση του αριθμού ακμών $e(t)$ και κόμβων $n(t)$, σε λογαριθμική-λογαριθμική κλίμακα, για τα εν ισχύ υπό-δίκτυα του Δικτύου Νομοθεσίας. Όλα τα υπό-δίκτυα ακολουθούν το νόμο δύναμης για την πύκνωση , με σταθερά καλή προσαρμογή. Τιμές της σταθεράς α: $1.142, 1.085, 1.167$ και 1.002, αντίστοιχα. Οι τιμές του συντελεστή προσδιορισμού (coefficient of determination) R^2 αποτυπώνονται επίσης. (Καλύτερη απεικόνιση έγχρωμα.)

4.3.3 Αξιολόγηση της σταθερότητας

Η έννοια του συστημικού κινδύνου εφαρμόζεται σε κάθε πολύπλοκο σύστημα και χιλιάδες κανονισμοί προσπαθούν να επισημοποιήσουν και να ρυθμίσουν τον συστημικό κίνδυνο σε διάφορους τομείς και κλάδους, π.χ., τραπεζικός συστημικός κίνδυνος. Η αλληλουχία αποτυχιών (cascading failures), σε ένα δίκτυο διασυνδεδεμένων συστημάτων έχει μελετηθεί στη βιβλιογραφία [34], αλλά έχει παραβλεφθεί στον ίδιο το νομικό τομέα. Όπως αναφέρεται στο [124] το νομικό σύστημα δεν πρέπει μόνο να προβλέπει συστημικές ελλείψεις στα συστήματα που έχει σχεδιαστεί για να ρυθμίσει, αλλά και να προβλέψει τον συστημικό κίνδυνο στο ίδιο το νομικό σύστημα. Με αφετηρία αυτή την ιδέα, σε αυτή την ενότητα, περιγράφουμε ένα πείραμα για την περαιτέρω μελέτη της ανθεκτικότητας του Δικτύου Νομοθεσίας. Αναγνωρίζουμε την

πολυπλοκότητα της αξιολόγησης αυτής που εξαρτάται από νομικούς εμπειρογνώμονες για την κατάλληλη ερμηνεία των νομικών συνεπειών των ακατέργαστων, από νομική σκοπιά, αποτελεσμάτων.

Ένα θεμελιώδες ζήτημα στην ανάλυση των πολύπλοκων δικτύων είναι η αξιολόγηση της σταθερότητας τους. Η αξιολόγηση αυτή έχει ως στόχο την κατανόηση και πρόβλεψη της συμπεριφοράς του συστήματος κάτω από κάθε είδους δυσλειτουργίες. Η ανθεκτικότητα αναφέρεται στην ικανότητα ενός δικτύου να συνεχίσει να λειτουργεί, όταν ένα κλάσμα των συστατικών του δυσλειτουγεί. Κατά τη διάρκεια των τελευταίων ετών έχει αξιολογηθεί ένας μεγάλος αριθμός δικτύων αναφορικά με την ανοχή τους σε σφάλματα και επιθέσεις, ενώ έχουν προταθεί διάφορες προσεγγίσεις του θέματος [6, 20, 103]. Συνήθως η ανοχή σε σφάλματα μετράται σε όρους μεταβολών της διαμέτρου ή του μεγέθους της γιγαντιαίας συνιστώσας των δικτύων υπό αξιολόγηση όταν ένα κλάσμα κόμβων αφαιρούνται με τυχαίο τρόπο. Αντίθετα, με την παραδοχή ότι ένας κακόβουλος πράκτορας θα στοχεύσει σκοπίμως στους πιο συνδεδεμένους κόμβους, η ανοχή σε επιθέσεις μετράται αφαιρώντας από το δίκτυο ένα κλάσμα από τους πιο συνδεδεμένους κόμβους, ταξινομημένους κατά φθίνουσα σειρά.

Η ύπαρξη ενός μοντέλου, όπως αυτό που παρουσιάζεται στο τρέχων κεφάλαιο, μας δίνει τη δυνατότητα να αναλύσουμε και να μετρήσουμε ποσοτικά τη συμπεριφορά του Δικτύου Νομοθεσίας, σε περίπτωση όπου κάποιοι από τους κόμβους του (νομικά έγγραφα) ή ακμές (συνδέσεις) μεταξύ των κόμβων αφαιρεθούν. Σε πραγματικές συνθήκες, αυτό μπορεί να συμβεί όταν θεσμοθετούνται νέοι νόμοι ή υφιστάμενοι νόμοι τροποποιούνται ή ακυρώνονται κατά τη διάρκεια μιας μεγάλης μεταρρυθμιστικής πρωτοβουλίας πχ, μεταρρυθμίσεις απορρύθμισης σε διάφορους τομείς της βιομηχανίας. Επιπλέον η συμμόρφωση με έναν νομικό κανόνα θα μπορούσε να απαιτήσει την εφαρμογή μέτρων που κάνουν την συμμόρφωση με άλλους κανόνες πιο δύσκολη. Ομοίως, επειδή οι νομικοί κανόνες είναι συχνά αλληλένδετοι και αλληλοεξαρτώμενοι μέσω τεχνικών, όπως οι νομικές παραπομπές και το δεδικασμένο, ο τρόπος που ένας κανόνας ερμηνεύεται και εφαρμόζεται θα μπορούσε να επηρεάσει την έννοια ή τη λειτουργία άλλων κανόνων [124].

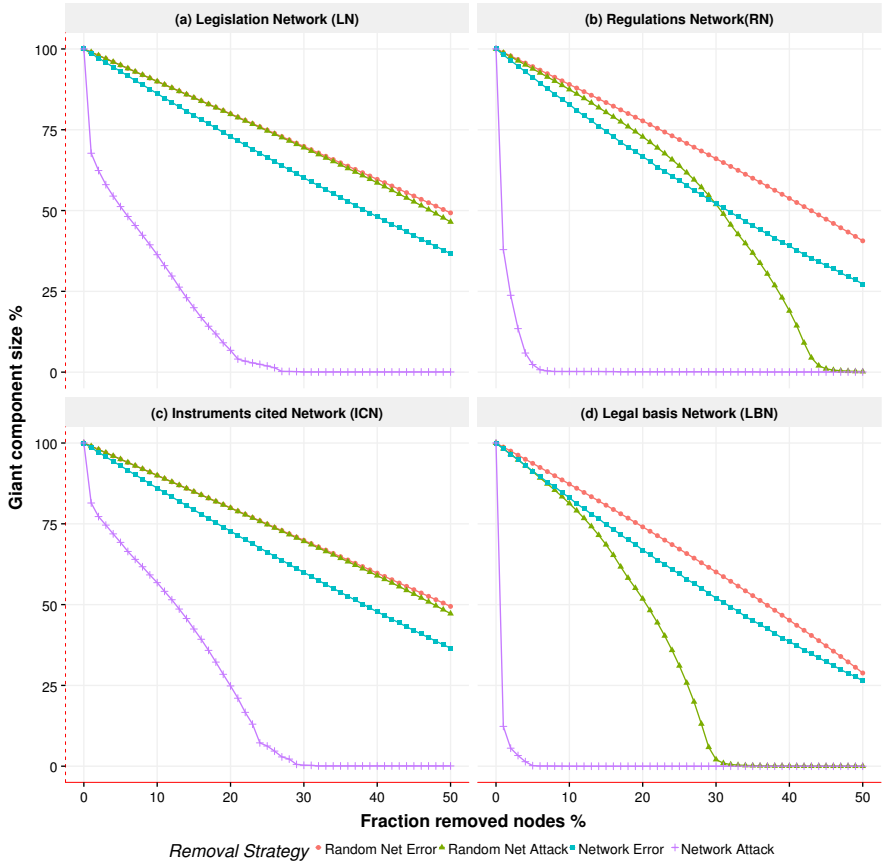
Αξιολογήσαμε τις αλλαγές στην γιγαντιαία συνιστώσα του δικτύου όταν ένα μικρό κλάσμα των κόμβων αφαιρείται. Η διαγραφή ενός κόμβου προκαλεί επίσης τη διαγραφή όλων των ακμών του. Επειδή οι εν λόγω αλλαγές, δεν μπορούν να εφαρμοστούν σε ανενεργή νομοθεσία, χρησιμοποιούμε στην ανάλυσή μας την τρέχουσα (ισχύουσα) έκδοση του κάθε υπό-δικτύου, όπως αυτά παρουσιάστηκαν στην Ενότητα 4.2.2, χρησιμοποιώντας τον Αλγόριθμο 4.3.

Προσομοιώσαμε 'λάθη' με τυχαία αφαίρεση κόμβων, ενώ για την προσομοίωση επιθέσεων αφαιρούμε κόμβους ανάλογα με το βαθμό κατανομής τους κατά φθίνουσα σειρά. Με τον τρόπο αυτό, οι επιθέσεις στοχεύουν στους υψηλά συνδεδεμένους κόμβους του δικτύου, αλλοιώνοντας δραστικά τη δομή σύνδεσης του δικτύου. Σε μελλοντική εργασία, σχεδιάζουμε να συμπεριλάβουμε ένα ευρύτερο φάσμα κριτηρίων για τον προσδιορισμό της σημασίας του αφαιρουμένου υπό επίθεση κόμβου, όπως Hits και PageRank [82].

Για καθένα από τα τέσσερα υπό-δίκτυα δημιουργήσαμε ένα τυχαίο, ισοδύναμο - με ίδιο αριθμό κόμβων και ακμών, τυχαίο δίκτυο Erdös-Rényi. Τα τυχαία αυτά δίκτυα μας βοηθούν στην οπτικοποίηση των επιδράσεων του φαινόμενου μικρού-κόσμου και νόμου δύναμης που

προηγούμενα πειραματικά αποδείξαμε ότι διέπουν το Δίκτυο Νομοθεσίας. Τα οκτώ υπό-δίκτυα αξιολογήθηκαν κάτω από τις υποθέσεις αποτυχίας / επίθεσης (error/ attack assumptions) με ένα ρυθμό απομάκρυνσης 5% των υπολειπόμενων κόμβων σε κάθε βήμα. Σε κάθε βήμα, υπολογίσαμε την γιγαντιαία συνιστώσα κάθε υπό-δικτύου σύμφωνα τους υπολειπόμενους κόμβους. Η όλη διαδικασία επαναλήφθηκε 1.000 φορές και υπολογίστηκαν οι μέσοι όροι του κλάσματος των κόμβων στην εναπομείνασα γιγαντιαία συνιστώσα.

Τα αποτελέσματα της αξιολόγησης σταθερότητας παρουσιάζονται στο Σχήμα 4.11. Για κάθε υπό-δίκτυο, απεικονίζεται το ποσοστό της γιγαντιαίας συνιστώσας σύμφωνα με το κλάσμα των κόμβων που αφαιρέθηκαν. Σε αντιστοιχία με την βιβλιογραφία περί δικτύων ελεύθερης-κλίμακας [5], το Δίκτυο Νομοθεσίας, παρουσιάζει εξαιρετική ανθεκτικότητα στις τυχαίες αποτυχίες. Ωστόσο, ως αποτέλεσμα αυτής της ανθεκτικότητας, στις περιπτώσεις σφαλμάτων στους κόμβους με τον υψηλότερο βαθμό κατανομής, το Δίκτυο Νομοθεσίας χάνει την συνοχή του και διασπάται νωρίτερα από τα ισοδύναμα τυχαία δίκτυα.



Σχήμα 4.11: Αξιολόγηση της σταθερότητας του Δικτύου Νομοθεσίας. Κλάσμα των κόμβων της γιγαντιαίας συνιστώσας ως συνάρτηση των κόμβων που έχουν αφαιρεθεί στα τέσσερα υπό-δίκτυα νομοθεσίας και τυχαία δίκτυα ίδιων διαστάσεων. Το υπό-δίκτυο Παραπομπών, Instruments cited (ICN), είναι το πιο ανθεκτικό από τα τέσσερα ενώ σε στοχευμένη επίθεση το υπό-δίκτυο Νομική Βάσης, Legal basis (LBN), είναι το λιγότερο ανθεκτικό. (Καλύτερη απεικόνιση έγχρωμα.)

Το Δίκτυο Νομοθεσίας (LN) είναι ένα δίκτυο ελεύθερης-κλίμακας, με πολλούς κόμβους χαμηλού βαθμού κατανομής και μερικούς με σημαντικά υψηλό βαθμό. Η τυχαία αφαίρεση κόμβων επηρεάζει κυρίως τους κόμβους χαμηλού βαθμού, αλλάζοντας έτσι οριακά την τοπολογία του δικτύου. Στις περιπτώσεις αυτές το δίκτυο συμπεριφέρεται όπως ένα τυχαίο δίκτυο. Ταυτόχρονα, η απομάκρυνση των υψηλά συνδεδεμένων κόμβων έχει μια καταστροφική επίδραση, αφήνοντας το δίκτυο άκρως διαιρεμένο.

Το υπό-δίκτυο Παραπομπών, Instruments cited (ICN), το οποίο προσομοιάζει ένα δίκτυο παραπομπών, παρουσιάζει την ανθεκτικότερη συμπεριφορά από όλα τα υπό αξιολόγηση υπόδίκτυα. Ωστόσο, χρησιμοποιώντας το σε τέτοια σενάρια θα οδηγηθούμε σε ανακριβή αποτελέσματα, διότι οι αναφορές τύπου 'Γενικές Παραπομπές' δεν μεταφέρουν καμία ειδική σημασία.

Από την άλλη πλευρά, το υπό-δίκτυο νομική βάση (Legal basis (LBN)) φαίνεται να είναι το λιγότερο ελαστικό από τα τέσσερα. Το υπό-δίκτυο αυτό αποτελείται μόνο από ακμές του τύπου *Νομική Βάση* και αντιπροσωπεύει την εσωτερική ιεραρχία της νομοθεσίας. Νόμοι με μεγάλη επιρροή, που χρησιμεύουν ως νομική βάση για πολλούς άλλους, διαδραματίζουν σημαντικό ρόλο στη νομοθεσία. Οι νόμοι αυτοί διατηρούν την συνοχή του δικτύου και τυχών τροποποιήσεις τους μπορεί να προκαλέσουν σοβαρές συνέπειες στο σώμα της νομοθεσίας.

Για ένα πραγματικό παράδειγμα, για τις συνέπειες που θα μπορούσε να επιφέρει μια και μόνο δικαστική απόφαση θεωρούμε την κάτωθι περίπτωση, Ireland's drug loophole case[2]. Το εφετείο της Ιρλανδίας, (Ireland's Court of Appeal), έκρινε τμήματα του, από το 1977, νόμου περί ναρκωτικών, (Misuse of Drugs Act) αντισυνταγματικά με το σκεπτικό ότι εγκρίθηκαν με υπουργική απόφαση και χωρίς διαβούλευση του ιρλανδικού κοινοβουλίου [142]. Καθώς ο νόμος ψηφίστηκε για πρώτη φορά, σχεδόν πριν από 40 χρόνια, διαδοχικές τροποποιήσεις προσέθεσαν ουσίες στην αρχική λίστα απαγορευμένων ουσιών. Με την απόφαση του δικαστηρίου όλες οι τροπολογίες κηρυχθήκαν άκυρες, ως εκ τούτου, ανοίγοντας ένα κενό στο νομικό σύστημα. Σύμφωνα με τα προηγούμενα ευρήματά μας σχετικά με τη δομή και την τοπολογία του δικτύου Νομοθεσίας, δίκτυο ελεύθερης-κλίμακας μικρού-κόσμου, αυξανόμενη διάμετρος και νόμος πύκνωσης, αναμένουμε ότι αυτού του είδους νομικά 'ατυχήματα' θα γίνουν πιο συχνά και με εκθετική αλληλουχία αποτυχιών.

Οι αρχές σε διάφορα επίπεδα εντός Ε.Ε π.χ., ευρωπαϊκό, εθνικό, τοπικό επίπεδο μπορούν να επωφεληθούν από την ανάλυση των επιπτώσεων των αλλαγών στο Δίκτυο Νομοθεσίας. Το προτεινόμενο μοντέλο μας μπορεί να παρέχει βοήθεια στους εμπειρογνώμονες, στο κομμάτι της αξιολόγησης των νομικών συνεπειών τυχών αλλαγών στην νομοθεσία.

4.4 Συμπεράσματα και Μελλοντικές επεκτάσεις

Η υιοθέτηση μιας οπτικής γωνίας δικτύου ήταν μία από τις πιο σημαντικές πρόσφατες εξελίξεις στις φυσικές και κοινωνικές επιστήμες, όπου πολύπλοκα συστήματα συχνά μοντελοποιούνται ως δίκτυα, προσφέροντας στους ερευνητές μια πολύ ισχυρή πολυεπίπεδη προοπτική. Στην παρούσα εργασία, εισάγουμε μία πρωτότυπη προσέγγιση για την μοντελοποίηση του δικαίου: το Δίκτυο Νομοθεσίας. Η εργασία αυτή διαφέρει σημαντικά από τις περισσότερες

[2] `http://edition.cnn.com/2015/03/11/europe/ireland-legal-drugs/`

προηγούμενες μελέτες νομικής αναλύσεως και από τη μονολιθική άποψη του σώματος της νομοθεσίας που προσφέρουν. Η προσέγγισή μας προσφέρει ένα πρότυπο για τη δημιουργία μιας συστηματικής εναλλακτικής δομής σε ένα φυσικά εξελισσόμενο κανονιστικό σύστημα. Το Δίκτυο Νομοθεσίας είναι ένα δίκτυο πολλαπλών σχέσεων που φιλοξενεί την ιεραρχία μεταξύ των πηγών του δικαίου και μπορεί να αντιπροσωπεύει σχέσεις διαφόρων κατηγοριών μεταξύ νομικών πηγών, μαζί με τη χρονική εξέλιξή τους.

Η εξέταση των δομικών ιδιοτήτων ενός δικτύου έχει θεμελιώδη σημασία για την κατανόηση της σύνθετης δυναμικής του υπό μοντελοποίηση συστήματος π.χ., εντοπισμός, αφανών οργανωτικών αρχών του σώματος της νομοθεσίας, ερμηνεία της επίδρασης της δομής του δικτύου σε μεμονωμένες νόμιμες πηγές και ποσοτικοποίηση της σχετικής σημασίας μιας νομικής πηγής μέσα σε ένα σώμα κειμένων. Η δομή του Δικτύου Νομοθεσίας είναι καλά διασυνδεδεμένη. Τα περισσότερα νομικά έγγραφα ανήκουν στη γιγαντιαία συνιστώσα και από οποιοδήποτε έγγραφο είναι δυνατό να προσεγγίσουμε σχεδόν κάθε άλλο, μόλις σε 7 βήματα, κατά μέσο όρο. Οι νομικές πηγές έχουν έντονη τάση να συνδέονται με νομικές πηγές του ιδίου τύπου, σχηματίζοντας ομάδες του ίδιου τύπου/τομέα. Η επικοινωνία μεταξύ των πολύ συσσωρευμένων περιοχών αραιά συνδεδεμένων κόμβων διατηρείται από μερικούς κόμβους, καθώς το Δίκτυο Νομοθεσίας είναι επίσης εξαιρετικά ετερογενές σε σχέση με τον αριθμό των συνδέσεων των νομικών πηγών. Η προέλευση αυτής της ετερογένειας, μπορεί να εξηγηθεί από τη διαδικασία της επιλεκτικής προσκόλλησης, η οποία ενισχύει τη δημοτικότητα των πηγών υψηλής κατάταξης. Η κατανομή βαθμών των νομικών εγγράφων ακολουθεί νόμο δύναμης (power law). Το δίκτυο αν και είναι ανθεκτικό στην τυχαία απώλεια κόμβων, είναι πολύ ευάλωτο σε επιθέσεις που στοχεύουν στα νομικά έγγραφα υψηλού βαθμού. Η διασύνδεση του Δικτύου Νομοθεσίας βασίζεται σε ένα μικρό αριθμό πολύ σημαντικών νομικών εγγράφων. Η τροποποίηση τέτοιων νομικών εγγράφων, όπως οι πράξεις τροποποίησης ή ακύρωσης, μπορεί να προκαλέσει μια χιονοστιβάδα απρόβλεπτων συνεπειών στο σώμα του νόμου. Τέλος, η χρονική εξέλιξη του ενεργού Δικτύου Νομοθεσίας αποκαλύπτει ότι γίνεται όλο και πιο πυκνό με την πάροδο του χρόνου, καθώς ο αριθμός των συνδέσεων μεταξύ των νομικών πηγών σε ισχύ αυξάνεται ταχύτερα από τον αριθμό των νομικών πηγών, ακολουθώντας έναν νόμο πύκνωσης.

Η εργασία μας παρέχει μια πρώτη προσέγγιση για την παροχή ενός μοντέλου για να εξηγήσουμε καλύτερα τη δομή και την εξέλιξη της νομοθεσίας. Η μαθηματική μοντελοποίηση των άφθονων συνδέσεων μεταξύ των νομικών πηγών ανοίγει νέες πόρτες έρευνας για τη βελτίωση της αποτελεσματικότητας του νομικού συστήματος με διάφορους τρόπους. Ο συγκερασμός των παραδοσιακών μεθόδων κατάταξης με τις μεθόδους ανάκτησης πληροφοριών στο διαδίκτυο [82], χρησιμοποιώντας το Δίκτυο Νομοθεσίας, αναμένεται να βελτιώσει την αποτελεσματικότητα των συστημάτων ανάκτησης νομικών πληροφοριών. Οι μέθοδοι διαφοροποίησης που βασίζονται σε γραφήματα μπορούν επίσης να προσφέρουν αξιόλογες βελτιώσεις όσον αφορά τον εμπλουτισμό των αποτελεσμάτων αναζήτησης , όπως παρουσιάζονται στο [73, 74] και αναλύονται διεξοδικά στο επόμενο κεφάλαιο 5. Υβριδικά συστήματα 'νομικών προτάσεων' (hybrid legal recommender systems), μπορεί να αναπτυχθούν αποτελεσματικότερα συνδυάζοντας το περιεχόμενο των νομικών πηγών και το συνεργατικό φιλτράρισμα βάσει χαρακτηριστικών του Δικτύου Νομοθεσίας.

Επιπρόσθετα, μέθοδοι ανίχνευσης κοινοτήτων σε γράφους [46] μπορούν να εφαρμοστούν στο Δίκτυο Νομοθεσίας για να αποκαλυφθεί η τομεακή δομή των νομικών πηγών και να εντοπιστούν ομάδες νομικών πηγών που μοιράζονται κοινές ιδιότητες και έχουν παρόμοιο ρόλο μέσα στο νομικό σώμα. Τεχνικές μοντελοποίησης θέματος (topic modelling), π.χ. αλγόριθμος λανθάνουσας κατανομής Latent dirichlet allocation [19] σε συνδυασμό με έναν αλγόριθμο εύρεσης διαδρομής, μπορεί επίσης να χρησιμοποιηθεί για την εξόρυξη των θεμάτων των νομικών πηγών και να προσδιορίσει τις σημασιολογικές ενώσεις τους με βάση τη δομή του δικτύου. Χρονικά προβλήματα 'παρασυρόμενων' θεμάτων (temporal topic drift) μπορούν επίσης να αντιμετωπιστούν εφαρμόζοντας τις προαναφερθείσες μεθοδολογίες στο Δίκτυο Νομοθεσίας.

Τεχνικές συσταδοποίησης (clustering techniques) μπορούν να χρησιμοποιηθούν για την ανίχνευση αποκλινόντων νομικών πηγών, ακόμη και σε επίπεδο παραγράφων, εάν οι νομικές πηγές έχουν μοντελοποιηθεί καταλλήλως και οι νομικές συνδέσεις έχουν ανιχνευθεί σε επίπεδο παραγράφων κειμένου, όπως αναλύεται διεξοδικά στο Κεφάλαιο 3. Τεχνικές ανάλυσης κανονιστικής συμμόρφωσης (analysis of regulatory compliance) και ανίχνευσης αντιφάσεων/συγκρούσεων μεταξύ των κανονισμών μπορούν επίσης να ωφεληθούν από την εξερεύνηση του Δικτύου Νομοθεσίας: σε μια προσέγγιση από πάνω προς τα κάτω, ξεκινώντας από το εάν δύο ή περισσότερες νομικές πηγές αναφέρονται σε άλλες νομικές πηγές, και καταλήγοντας σε πιο λεπτομερές επίπεδο αποκαλύπτοντας το ακριβές τμήμα αναφοράς και το είδος της σχέσης που συνδέει τις νομικές πηγές. Οι μέθοδοι περίληψης, μίας ή περισσότερων νομικών πηγών, μπορούν επίσης να επωφεληθούν από τις σχέσεις μεταξύ νομικών πηγών που προσδιορίζονται στο Δίκτυο Νομοθεσίας.

Παράλληλα με τα προηγούμενα, το μοντέλο που προτείνεται μπορεί να αξιοποιηθεί για την γραφική απεικόνιση του δικαίου. Ένα σύστημα απεικόνισης του Δικτύου Νομοθεσίας μπορεί να βοηθήσει τόσο τους πολίτες όσο και τους νομικούς εμπειρογνώμονες, βοηθώντας τους να πλοηγηθούν εύκολα στο σώμα του νόμου, επισημαίνοντας πρότυπα, αποκαλύπτοντας συστάδες και συναφείς συνδέσεις, αποκαλύπτοντας επικαλύψεις και πιθανές συγκρούσεις. Ένα άλλο πλεονέκτημα μιας τέτοιας προσέγγισης έγκειται στο γεγονός ότι το σώμα νομοθετικών κειμένων μπορεί να αξιοποιηθεί όχι μόνο από την παραδοσιακή οπτική γωνία, αλλά και ως γράφημα υπερ-κειμενικών πληροφοριών με χρονικές ιδιότητες. Για παράδειγμα, θα είναι ευκολότερο για τους νομοθέτες να παρακολουθήσουν την επίδραση μιας πιθανής αλλαγής στο σύνολο του νομικού πλαισίου. Σε ένα τέτοιο σύστημα, η χρήση οντολογιών και τεχνολογιών συνδεδεμένων δεδομένων θα εμπλουτίσουν περαιτέρω την προστιθέμενη αξία του Δικτύου Νομοθεσίας.

Κεφάλαιο 5

Διαφοροποιημένη Ανάκτηση Πληροφορίας σε Νομικές Πηγές

Στο κεφάλαιο αυτό μελετάμε την διαφοροποιημένη ανάκτηση νομικών πηγών. Αρχικά καταδεικνύουμε την ανάγκη επέκτασης των τεχνικών διαφοροποίησης, ειδικά για το σενάριο της διαφοροποιημένης ανάκτησης νομικών κειμένων και οριοθετούμε το πρόβλημα. Στην συνέχεια εισάγουμε με βάση τα ιδιαίτερα χαρακτηριστικά των νομικών κειμένων εξειδικευμένα κριτήρια διαφοροποίησης. Παράλληλα, προσαρμόζουμε στο συγκεκριμένο πρόβλημα και στα κριτήρια που εισάγουμε διαδεδομένους αλγόριθμους που έχουν προταθεί για την κάλυψη διαφορετικών αναγκών π.χ., την δημιουργία περιλήψεων, την διαφοροποιημένη κατάταξη σε γράφους, παράλληλα με αλγόριθμους διαφοροποίησης αποτελεσμάτων αναζήτησης. Τέλος, πραγματοποιούμε εκτενή πειραματική αξιολόγηση των προαναφερθέντων μεθόδων και κριτηρίων διαφοροποίησης σε ποικίλες περιπτώσεις, με πραγματικές συλλογές νομικών εγγράφων, από διαφορετικά νομικά συστήματα, χρησιμοποιώντας διεθνώς αποδεκτές μετρικές και αντικειμενική μεθοδολογία επισημείωσης του συνόλου δεδομένων, παρέχοντας όρια εξισορρόπησης μεταξύ της ενίσχυσης των σχετικών εγγράφων ή της ποικολομορφίας του συνόλου αποτελεσμάτων [72, 73, 74].

5.1 Κίνητρο και Συνεισφορά

Στην σημερινή εποχή, ως συνέπεια των πρωτοβουλιών ανοιχτών δεδομένων, παρατηρείται μια τεράστια αύξηση στον αριθμό των συνόλων νομικών δεδομένων που είναι ελεύθερα διαθέσιμα. Νομικά δεδομένα και κείμενα που ήταν προηγουμένως διαθέσιμα μόνο σε ένα εξειδικευμένο κοινό μορφή είναι τώρα ελεύθερα διαθέσιμα στο διαδίκτυο.

Διαδικτυακές πύλες, portals, όπως το EUR-Lex[1], United States Code[2], United Kingdom[3], Brazil[4] και Australian[5], παραθέτοντας ορισμένες, παρέχουν πρόσβαση σε εκατομμύρια κανονισμούς, δικαστικές υποθέσεις ή διοικητικές αποφάσεις. Τέτοιες πύλες παρέχουν πολλα-

[1] http://eur-lex.europa.eu/
[2] http://uscode.house.gov/
[3] http://www.legislation.gov.uk/
[4] http://www.lexml.gov.br/
[5] https://www.comlaw.gov.au/

πλές δυνατότητες αναζήτησης, έτσι ώστε να βοηθήσουν τους τελικούς χρήστες να ανακτήσουν τις πληροφορίες που χρειάζονται. Για παράδειγμα, ο χρήστης μπορεί να εκτελέσει απλές εργασίες αναζήτησης ή χρησιμοποιώντας προκαθορισμένα κριτήρια, π.χ., χρόνο, νομική βάση, θεματική κατηγορία, να ανακτήσει σχετικά με την πληροφοριακή του ανάγκη νομικά έγγραφα.

Ταυτόχρονα, το νομικό πεδίο δημιουργεί ένα τεράστιο και συνεχώς αυξανόμενο πλήθος ανοιχτών δεδομένων, αξιοποιώντας έτσι το επίπεδο συνειδητοποίησης της νομοθεσίας των ενδιαφερομένων. Δικαστικές αποφάσεις, δικαστικά προηγούμενα και ερμηνείες των νόμων, δημιουργούν μια μεγάλη δεξαμενή δεδομένων με σχετικές και χρήσιμες νομικές πηγές, στον οποίο συνήθως παραμένουν κρυμμένα γεγονότα και ιδέες που θα μπορούσαν να βοηθήσουν να στηριχτεί ένα νομικό επιχείρημα.

Ας σκεφτούμε, για παράδειγμα, την περίπτωση ενός δικηγόρου με ειδίκευση σε διπλώματα ευρεσιτεχνιών, ο οποίος θέλει να βρει σχετικά διπλώματα ευρεσιτεχνίας ως περίπτωση αναφοράς και υποβάλλει ένα ερώτημα χρήστη για την ανάκτηση των πληροφοριών. Ένα διαφοροποιημένο αποτέλεσμα, δηλαδή, ένα αποτέλεσμα που περιέχει αρκετές αξιώσεις, με ετερογενής κανονιστικές απαιτήσεις/ συμβάσεις - με διαφορετικό αριθμό εφευρετών και λοιπών χαρακτηριστικών - είναι διαισθητικά πιο κατατοπιστικό από ένα ομοιογενές σύνολο αποτελεσμάτων που περιέχουν μόνο τα διπλώματα ευρεσιτεχνίας με παρόμοια χαρακτηριστικά. Στο κεφάλαιο αυτό, προτείνουμε νέες μεθόδους για τον αποδοτικό και αποτελεσματικό χειρισμό παρόμοιων προκλήσεων κατά την αναζήτηση πληροφοριών σε νομικά κείμενα.

Η διαφοροποίηση είναι μια μέθοδος που στοχεύει στην βελτίωση της ικανοποίησης του χρήστη, αυξάνοντας την ποικιλία των πληροφοριών που εμφανίζονται σε αυτόν. Κατά συνέπεια, ο αριθμός των επαναλαμβανόμενων στοιχείων σε μια λίστα αποτελεσμάτων αναζήτησης θα πρέπει να μειωθεί, ενώ η πιθανότητα ότι ο χρήστης θα μείνει ικανοποιημένος με οποιαδήποτε από τα εμφανιζόμενα αποτελέσματα θα αυξηθεί. Υπάρχει εκτεταμένη βιβλιογραφική εργασία σχετικά με τεχνικές διαφοροποίησης (βλέπε Ενότητα 2.1.2), όπου η βασική ιδέα είναι η επιλογή ενός μικρού σύνολου των αποτελεσμάτων που είναι αρκετά ανόμοια, σύμφωνα με ένα κατάλληλο μέτρο ομοιότητάς.

Η εφαρμογή τεχνικών διαφοροποίησης στα νομικά συστήματα πληροφοριών μπορεί να είναι χρήσιμη όχι μόνο για τους πολίτες αλλά και για τους νομοθέτες καθώς και λοιπούς ενδιαφερόμενους π.χ., επιχειρήσεις και μεγάλους οργανισμούς. Έχοντας μια μεγάλη εικόνα των διαφοροποιημένων αποτελέσματων, οι ενδιαφερόμενοι μπορούν να επιλέξουν ή να προσαρμόσουν κατάλληλα το νομικό καθεστώς που ταιριάζει καλύτερα στις επιχειρησιακές τους ανάγκες, βοηθώντας τους έτσι να λειτουργήσουν αποτελεσματικότερα. Επιπλέον, οι τεχνικές αυτές μπορούν επίσης να βοηθήσουν τους νομοθέτες, δεδομένου ότι η βαθιά κατανόηση των νομικών διαφοροποίησων προωθεί την εξέλιξη σε καλύτερες και δικαιότερες νομοθετικές ρυθμίσεις για την κοινωνία [7].

Η εφαρμογή τεχνικών διαφοροποίησης στα νομικά συστήματα πληροφοριών έχει αρκετές ομοιότητες με τη διαφοροποίηση αποτελεσμάτων αναζήτησης, για παράδειγμα, το γεγονός ότι, και στις δύο περιπτώσεις, τα στοιχεία προς διαφοροποίηση έχουν κειμενικές περιγραφές. Παρόλα αυτά, υπάρχουν και ουσιώδεις διαφορές που επιβάλλουν την ανάγκη ανάλυσης και επέκτασης/προσαρμογής των αλγορίθμων/κριτηρίων διαφοροποίησης, ειδικά για το σενάριο

της διαφοροποίησης νομικών πηγών.

Αναλυτικά, προσαρμόζουμε στο συγκεκριμένο πρόβλημα, διαφοροποίησης αποτελεσμάτων αναζήτησης σε νομικά δεδομένα, μεθόδους που έχουν προταθεί στην βιβλιογραφία για την κάλυψη ετερογενών αναγκών. Πιο συγκεκριμένα, προσαρμόζουμε και αξιολογούμε την απόδοση αλγορίθμων που χρησιμοποιούνται για: α) την δημιουργία περιλήψεων κειμένων [LexRank [39] και Biased LexRank [106]], β) την διαφοροποιημένη κατάταξη σε γράφους (graph-based ranking) [DivRank [95] και Grasshopper [159]] και γ) την διαφοροποίηση αποτελεσμάτων αναζήτησης [MMR [25], Max-sum [58], Max-min [58] και Mono-objective

Επιπρόσθετα εξετάζοντας την φύση των νομικών πηγών, εισάγουμε εξειδικευμένα κριτήρια στις παραπάνω μεθόδους, αναλύοντας ταυτόχρονα την επίδραση διαφόρων χαρακτηριστικών των νομικών πηγών στον υπολογισμό της ομοιότητας των εγγράφων με το ερώτημα του χρήστη και των εγγράφων μεταξύ τους.

Παράλληλα, αξιολογούμε την απόδοση των προαναφερθέντων μεθόδων και κριτηρίων διαφοροποίησης σε ποικίλες περιπτώσεις (σενάρια διαφοροποίησης), χρησιμοποιώντας πραγματικά δεδομένα, από διαφορετικές κατηγορίες τύπων νομικών συστημάτων. Η αξιολόγηση μας στηρίζεται σε διεθνώς αποδεκτές μετρικές και αντικειμενικές μεθόδους επισημείωσης του συνόλου δεδομένων.

Τα ευρήματά μας αποκαλύπτουν ότι:

- η υιοθέτηση μεθόδων διαφοροποίησης, στο πλαίσιο ανάκτησης νομικής πληροφορίας, επιφέρει στατιστικά αξιοσημείωτες βελτιώσεις όσον αφορά τον εμπλουτισμό των αποτελεσμάτων αναζήτησης με κατά τα άλλα κρυφές πτυχές του νομικού χώρου γύρω από το ερώτημα του χρήστη

- τα κριτήρια διαφοροποίησης που προτείνουμε επίσης παρέχουν στατιστικά αξιοσημείωτα διαφοροποιημένα υποσύνολα ανακτημένων νομικών εγγράφων

- οι μέθοδοι διαφοροποίησης αποτελεσμάτων αναζήτησης ξεπερνούν, στο πλαίσιο της νομικής διαφοροποίησης, άλλες προσεγγίσεις, π.χ. μέθοδοι δημιουργίας περιλήψεων κειμένων ενώ οι μέθοδοι κατάταξης σε γράφους επιφέρουν αξιοσημείωτα αποτελέσματα μόνο σε περιπτώσεις πληρότητας του σώματος νομικών πηγών.

- η ανάλυση ακρίβειας που διεξάγουμε προσφέρει όρια εξισορρόπησης για τα συστήματα ανάκτησης νομικής πληροφορίας, που επιθυμούν να ισορροπήσουν μεταξύ της ενίσχυσης των σχετικών εγγράφων, (ομοιότητα αποτελεσμάτων) , ή να δειγματοληπτήσουν τον χώρο νομικής πληροφορίας γύρω από το ερώτημα (ποικιλομορφία αποτελεσμάτων).

5.2 Διαδικασία Διαφοροποίησης Νομικών Κειμένων

Αρχικώς ορίζουμε το πρόβλημα που μας απασχολεί και παρέχουμε μια επισκόπηση της διαδικασίας διαφοροποίησης. Στη συνέχεια, παρουσιάζουμε αναλυτικά τα προτεινόμενα κριτήρια διαφοροποίησης και περιγράφουμε τους αλγόριθμους διαφοροποίησης που εξετάζουμε.

5.2.1 Ορισμός Προβλήματος

Η διαφοροποίηση των αποτελεσμάτων αναζήτησης είναι ένα συμβιβασμός, εξισορρόπηση, μεταξύ της εύρεσης σχετικών με το ερώτημα του χρήστη εγγράφων και της ποικιλομορφίας του συνόλου των αποτελεσμάτων. Λαμβάνοντας υπόψη ένα σύνολο νομικών εγγράφων και ένα ερώτημα, στόχος μας είναι να βρούμε ένα σύνολο σχετικών και αντιπροσωπευτικών με το ερώτημα του χρήστη εγγράφων και να επιλέξουμε αυτά τα έγγραφα κατά τέτοιο τρόπο, ώστε η ποικιλομορφία του συνόλου να μεγιστοποιείται. Πιο συγκεκριμένα, το πρόβλημα μπορεί να οριστεί ως εξής:

Ορισμός 5.1 (Διαφοροποίηση Νομικών Εγγράφων)**.** *Έστω q ένα ερώτημα του χρήστη και N ένα σύνολο από νομικά έγγραφα, σχετικά με το ερώτημα του χρήστη. Βρείτε ένα υποσύνολο $S \subset N$ εγγράφων που μεγιστοποιεί μία συνάρτηση στόχο f η οποία ποσοτικοποιεί την ποικιλομορφία / ετερογένεια των εγγράφων στο S.*

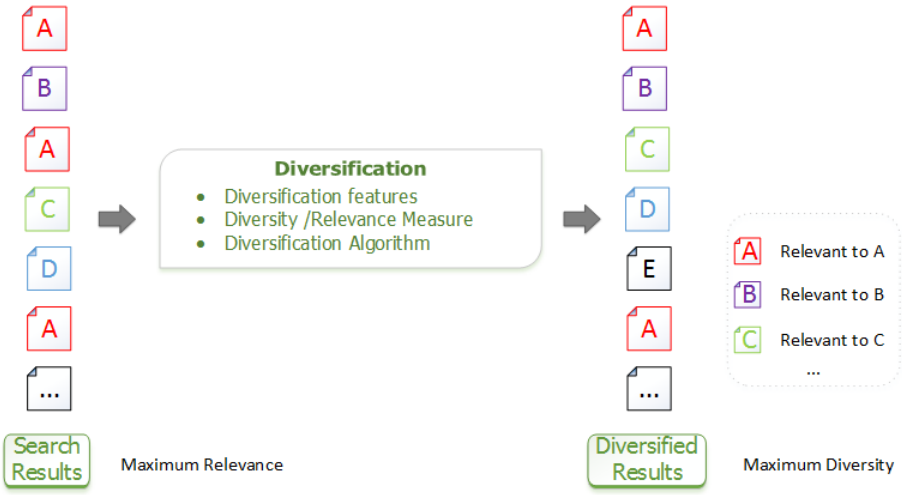
Το Σχήμα 5.1, απεικονίζει τη συνολική ροή εργασίας της διαδικασίας διαφοροποίησης. Αρχικά, ο χρήστης εκφράζει την πληροφοριακή του ανάγκη μέσω ενός ερωτήματος και ανακτά σχετικά με την πληροφοριακή του ανάγκη έγγραφα, όπως φαίνεται στην αριστερή στήλη του Σχήματος 5.1, όπου διαφορετικά ονόματα/ χρώματα αντιπροσωπεύουν βασικές πτυχές/ θέματα των εγγράφων. Από την αρχική κατάταξη των εγγράφων, με κριτήριο την συνάφεια με το ερώτημα, έχοντας ως στόχο την ταυτόχρονη επίτευξη συνάφειας και ποικιλομορφίας, παράγουμε μια διαφοροποιημένη κατάταξη των εγγράφων. Η λίστα αυτή απεικονίζεται στην δεξιά στήλη του Σχήματος 5.1. Σημειώνεται ότι το έγγραφο E, ήταν αρχικά κρυμμένο στην αριστερή στήλη που απεικονίζει την αρχική κατάταξη των εγγράφων με βάση την συνάφεια με το ερώτημα.

Για την εφαρμογή της μεθόδου θα πρέπει να οριστούν:

- **Κριτήρια Διαφοροποίησης**, χαρακτηριστικά των νομικών εγγράφων που θα χρησιμοποιηθούν στη διαδικασία διαφοροποίησης.
- **Συναρτήσεις αποστάσεων**, ορίζουμε συναρτήσεις μέτρησης της απόστασης μεταξύ δύο διανυσμάτων χαρακτηριστικών. Σημειώνουμε ότι στην παρούσα εργασία, όπως και γενικότερα σε θέματα ανάκτησης πληροφοριών [60], ο όρος ¨απόσταση' χρησιμοποιείται ανεπίσημα ως ένα μέτρο ανομοιότητας με βάση τα χαρακτηριστικά των εξεταζομένων εγγράφων.
- **Αλγόριθμοι Διαφοροποίησης**, εφαρμόζουμε ευριστικούς αλγόριθμους διαφοροποίησης στις καταχωρήσεις χρηστών.

5.2.2 Κριτήρια Διαφοροποίησης

Στο μοντέλο διανυσματικού χώρου (vector space model) [125], κάθε έγγραφο u μπορεί να αναπαρασταθεί ως διάνυσμα όρων $U = (is_{w1u}, is_{w2u}, ..., Is_{wmu})^T$, όπου $w_1, w_2, ..., w_m$ είναι όλοι οι διαθέσιμοι όροι (terms) , και is μπορεί να είναι οποιοδήποτε δημοφιλές σχήμα



Σχήμα 5.1: Επισκόπηση Διαφοροποίησης Νομικών Εγγράφων. Κάθε έγγραφο αναπαρίσταται με διαφορετικά ονόματα/ χρώματα που αντιπροσωπεύουν βασικές πτυχές/ θέματα των εγγράφων. Η διαφοροποιημένη κατάταξη των εγγράφων στην δεξιά λίστα, προκύπτει εφαρμόζοντας τεχνικές διαφοροποίησης στην αρχική κατάταξη, αριστερή λίστα των εγγράφων, με κριτήριο την συνάφεια με το ερώτημα.(Καλύτερη απεικόνιση έγχρωμα.)

ευρετηρίασης (indexing schema), π.χ., $tf, tf - idf, logtf - idf$. Τα ερωτήματα των χρηστών αναπαριστώνται με τον ίδιο τρόπο όπως τα έγγραφα.

Συνήθως οι τεχνικές διαφοροποίησης μετρούν την ποικιλομορφία όσον αφορά την κειμενική περιγραφή, όπου μόνο η ομοιότητα κειμένου μεταξύ των στοιχείων χρησιμοποιείται για την ποσοτικοποίηση της ομοιότητας των πληροφοριών. Στην ενότητα αυτή, επεκτείνουμε την έννοια της διαφορετικότητας σε επικουρικά χαρακτηριστικά/ διαστάσεις, εκτός από την ομοιότητα του κειμένου. Για να εντοπισουμε αυτές τις διαστάσεις εξετάζουμε τα μοναδικά χαρακτηριστικά των νομικών εγγράφων. Τα Νομικά έγγραφα έχουν κάποια αξιοσημείωτα χαρακτηριστικά, όπως ότι είναι εγγενώς πολυ-θεματικά, ακολουθούν μια 'αυστηρά' δομημένη γλώσσα και καλύπτουν ένα ευρύ και άνισα κατανεμημένο πεδίο νομικών θεμάτων [89].

Με βάση τα προαναφερθέντα χαρακτηριστικά των νομικών εγγράφων ορίζουμε τα ακόλουθα κριτήρια διαφοροποίησης:

– **Θεματικές Κατηγορίες**. Η με βάση το κύριο θέμα ενός εγγράφου ομαδοποίηση συσχετιζόμενων/ συναφών εγγράφων αποτελεί ιδιαίτερα διαδεδομένη τεχνική ταξινόμησης νομικών εγγράφων. Επιθυμούμε την επιλογή θεματικών κατηγοριών που καλύπτουν διαφορετικές ερμηνείες σε σχέση με τις ανάγκες πληροφόρησης των χριστών. Η θεματική ομοιότητα δύο εγγράφων που έχουν θεματικά σύνολα X_u και X_v υπολογίζεται με βάση την ομοιότητα Jacard ως εξής:

$$S_x(u, v) = \frac{|X_u \cap X_v|}{|X_u \cup X_v|} \tag{5.1}$$

– **Χρόνος**. Ο χρόνος αποτελεί μια σημαντική παράμετρο διαφοροποίησης, καθώς στις περισ-

σότερες περιπτώσεις θέματα και υποθέματα συσχετιζόμενα με τα ερωτήματα των χρηστών είναι χρονικά διφορούμενα, εξαιτίας της δυναμικής εξέλιξης και των εξαρτήσεων στο νομικό σύστημα. Η χρονική ομοιότητα μεταξύ των εγγράφων u και v, που έχουν χρονικές σφραγίδες t_u και t_v υπολογίζεται βάση της διαφοράς των κανονικοποιημένων χρονικών σφραγίδων τους με Min-Max κανονικοποίηση ως εξής:

$$S_t(u, v) = 1 - |t_{norm}(u) - t_{norm}(v)| \tag{5.2}$$

– **Ποιότητα γραφής (Readability)**. Η ποιότητα γραφής είναι ένας παράγοντας διαφοροποίησης, δεδομένου ότι εκφράζει το βαθμό κατανόησης του κειμένου. Έτσι, είναι σημαντικό ένα σύνολο αποτελεσμάτων να περιλαμβάνει διαφορετικά επίπεδα καταληπτότητας - αναγνωσιμότητας. Η αναγνωσιμότητα ενός κειμένου υποδηλώνει το επίπεδο δυσκολίας του γραπτού κειμένου και προκύπτει από την εφαρμογή ενός τύπου αναγνωσιμότητας που λαμβάνει υπόψη ποιοτικά χαρακτηριστικά ενός κειμένου, όπως μήκος λέξεων και προτάσεων. Η πιο διαδεδομένη από αυτές τις φόρμουλες είναι η Flesch Reading Ease Score[6], η οποία συνδυάζει μέσο αριθμό συλλαβών ανά λέξη και μέσο μήκος πρότασης για να παράγει μία βαθμολογία καταληπτότητας. Συγκεκριμένα, παράγει ένα σκορ στο διάστημα [0,100], με υψηλότερες τιμές να υποδεικνύουν ευκολότερα κείμενα. Η ομοιότητα Ποιότητας γραφής μεταξύ των εγγράφων u και v, που έχουν αριθμητικές τιμές Ποιότητας γραφής r_u και r_v υπολογίζεται βάση της διαφοράς της κανονικοποιημένων τιμών Ποιότητας γραφής τους με Min-Max κανονικοποίηση ως εξής:

$$S_r(u, v) = 1 - |r_{norm}(u) - r_{norm}(v)| \tag{5.3}$$

– **Κειμενικό περιεχόμενο**. Διάφορες γνωστές λειτουργίες από τη βιβλιογραφία (π.χ. Jaccard, ομοιότητα συνημίτονου κ.λπ.) μπορούν να χρησιμοποιηθούν στον υπολογισμό της ομοιότητας κειμένου των νομικών εγγράφων. Επιλέγοντας την συνάρτηση συνημίτονου ομοιότητα ως μέτρο ομοιότητας, η ομοιότητα του κειμένου μεταξύ εγγράφων u και v, με διανύσματα όρων (term vectors) U και V είναι:

$$S_c(u, v) = \cos(u, v) = \frac{U \cdot V}{\| U \| \| V \|} \tag{5.4}$$

5.2.3 Συναρτήσεις Αποστάσεων

Στην Ενότητα 5.2.2 περιγράφηκαν οι υλοποιήσεις των κριτηρίων διαφοροποίησης, μέσω διανυσμάτων χαρακτηριστικών που αντιπροσωπεύουν διάφορες εκφάνσεις των νομικών πηγών. Οι αλγόριθμοι διαφοροποίησης χρησιμοποιούν αυτά τα διανύσματα για να υπολογίζουν σε κάθε βήμα ένα αθροιστικό σκορ διαφοροποίησης για κάθε έγγραφο. Προκειμένου να παράγουμε σκορ διαφοροποίησης, πρέπει να ορίσουμε συναρτήσεις για την ομοιότητα, απόσταση δύο εγγράφων καθώς και την ομοιότητα ενός εγγράφου με το ερώτημα του χρήστη.

[6] `http://en.wikipedia.org/wiki/Readability`

– Ομοιότητα Εγγράφων: Διάφορες προταθείσες στην βιβλιογραφία μέθοδοι μπορούν να υιοθετηθούν για τον υπολογισμό της ομοιότητας μεταξύ δύο εγγράφων. Στη παρούσα εργασία επιλέγουμε την ευρέως χρησιμοποιούμενη συνάρτηση ομοιότητας συνημίτονου (cosine similarity function) και ορίζουμε την ομοιότητα μεταξύ δύο εγγράφων, u και v, με διανύσματα όρων (term vectors) U και V ως εξής:

$$sim(u, v) = \cos(u, v) = \frac{U \cdot V}{\| U \| \| V \|} \tag{5.5}$$

ή γενικεύοντας, με χρήση των προτεινόμενων κριτηρίων διαφοροποίησης, η ομοιότητα δύο εγγράφων u, v υπολογίζεται ως γραμμική σταθμισμένη συνάρτηση των επί μέρους τιμών ομοιοτήτων Θεματικής Κατηγορίας, Χρόνου, Ποιότητας γραφής και Κειμενικού περιεχομένου ως εξής:

$$sim(u, v) = \sum_{i=1}^{|4|} w_i \; feat_i(u, v) = w_1 \; S_c(u, v) + w_2 \; S_x(u, v) + w_3 \; S_t(u, v) + w_4 \; S_r(u, v) \tag{5.6}$$

με βάρη $\sum_{i=1}^{|4|} w_i = 1$.

– Απόσταση Εγγράφων: Η απόσταση δύο εγγράφων, u και v είναι :

$$d(u, v) = 1 - sim(u, v) \tag{5.7}$$

– Ομοιότητα εγγράφου με το ερώτημα του χρήστη. Η ομοιότητα του ερωτήματος του χρήστη, q, με ένα έγγραφο, u, μπορεί να εξαχθεί από το αρχικό σκορ ταξινόμησης του IR συστήματος, ή να υπολογιστεί χρησιμοποιώντας την ίδια συνάρτηση όπως και για την προαναφερθείσα περίπτωση ομοιότητας εγγράφων πχ συνάρτηση ομοιότητας συνημίτονου στα αντίστοιχα διανύσματα όρων:

$$r(q, u) = \cos(q, u) \tag{5.8}$$

5.2.4 Αλγόριθμοι Διαφοροποίησης

Οι μέθοδοι διαφοροποίησης χρησιμοποιούν μια ταξινομημένη με βάση την συνάφεια με το ερώτημα λίστα εγγράφων και επανα-κατατάσσουν τα έγγραφα, έτσι ώστε τα αρχικά k έγγραφα να καλύπτουν περισσότερες υποκατηγορίες σχετικές με το αρχικό ερώτημα. Δεδομένου ότι το πρόβλημα της εύρεσης ενός βέλτιστου συνόλου διαφοροποιημένων εγγράφων είναι NP-complete [40], ένας άπληστος αλγόριθμος χρησιμοποιείται συχνά για την επαναληπτική επιλογή του διαφοροποιημένου συνόλου S.

Έστω N ένα σύνολο από νομικά έγγραφα σχετικά με το ερώτημα q του χρήστη, νομικά έγγραφα $u, v \in N$, $r(q, u)$ ο βαθμός ομοιότητας του u με το ερώτημα q, $d(u, v)$ είναι η απόσταση μεταξύ των εγγράφων u και v, S, το σύνολο διαφοροποιημένων εγγράφων και

$\lambda \in [0..1]$ παράμετρος που προσδιορίζει το συμβιβασμό (trade-off) μεταξύ συνάφειας και ανομοιότητας. Στην παρούσα εργασία, εστιάζουμε στις κάτωθι αντιπροσωπευτικές κατηγορίες μεθόδων διαφοροποίησης , που έχουν προταθεί σε προηγούμενες εργασίες:

Αλγόριθμοι Διαφοροποίησης Αποτελεσμάτων Αναζήτησης

– **MMR:** Η μέθοδος Maximal Marginal Relevance [25], είναι μια άπληστη μέθοδος που συνδυάζει την συνάφεια με το ερώτημα και την νέα πληροφορία (information novelty). Κατασκευάζει με επαναληπτική διαδικασία το σύνολο διαφοροποιημένων εγγράφων S, επιλέγοντας σε κάθε βήμα τα έγγραφα που μεγιστοποιούν την κάτωθι συνάρτηση στόχο:

$$f_{MMR}(u,q) = (1-\lambda)\ \ r(u,q) + \lambda \sum_{v \in S} d(u,v) \tag{5.9}$$

με τον έναν παράγοντα να υπολογίζει την ομοιότητα μεταξύ των εγγράφων ενώ ο άλλος την ομοιότητα μεταξύ του κάθε εγγράφου και του ερωτήματος. Επιπλέον, μια παράμετρος λ ελέγχει το βαθμό συμβιβασμού, εξισορροπώντας τους δύο παράγοντες.

Αλγόριθμος 5.1 Αλγόριθμος Διαφοροποίησης MMR

Input: Set of candidate results N, size of diverse set k
Output: Set of diverse results $S \subseteq N, |S| = k$
  $S = \emptyset$
  $N_i = argmax_{N v \in N}(r(v,q))$ ▷ initialize with the highest relevant to the query document
  Set $S = S \cup \{i\}$
  Set $N = N \setminus \{i\}$
  while $|S| < k$ **do**
    Find $u = argmax_{N v \in N}(f_{MMR}(v,q))$ ▷ iter. select document that maximize Eq.5.9
    Set $S = S \cup \{u\}$
    Set $N = N \setminus \{u\}$
  end while

Η μέθοδος MMR κατατάσσει τα έγγραφα με κριτήριο την συνάφεια τους με το ερώτημα για $\lambda = 0$ (αρχική κατάταξη των εγγράφων), ενώ για $\lambda = 1$ κατατάσσει τα έγγραφα με κριτήριο την μέγιστη ποικιλομορφία. Για ενδιάμεσες τιμές του $\lambda \in [0..1]$, βελτιστοποιείται ένας γραμμικός συνδυασμός των δύο κριτηρίων. Ο Αλγόριθμος MMR 5.1, αρχικοποιεί το σύνολο S με το έγγραφο που έχει τη μεγαλύτερη συνάφεια με το ερώτημα.

– **Max-sum:** Η μέθοδος Max-sum [58] στοχεύει στη μεγιστοποίηση της συνάφειας και της διαφορετικότητας στο τελικό σύνολο αποτελεσμάτων. Αυτό επιτυγχάνεται με μια άπληστη προσέγγιση, Αλγόριθμος 5.2, που επιλέγει, σε κάθε επανάληψη, ένα ζεύγος εγγράφων που μεγιστοποιεί την Εξίσωση 5.10 .

$$f_{MAXSUM}(u,v,q) = (1-\lambda)\ (r(u,q) + r(v,q)) + 2\lambda\ d(u,v) \tag{5.10}$$

όπου (u, v) είναι ένα ζεύγος εγγράφων, δεδομένου ότι η μέθοδος εξετάζει ζεύγη έγγραφων για εισαγωγή. Όταν το $|S|$ είναι περιττό, στην τελική φάση του αλγορίθμου ένα έγγραφο N επιλέγεται αυθαίρετα για να εισαχθεί στο σύνολο των διαφοροποιημένων αποτελεσμάτων S.

Ο Αλγόριθμος Max-sum 5.2, σε κάθε βήμα, εξετάζει τα ζεύγη αποστάσεων των υποψήφιων εγγράφων N και επιλέγει το ζεύγος με τη μέγιστη απόσταση για να εισάγει στο σύνολο διαφοροποιημένων αποτελεσμάτων S.

Αλγόριθμος 5.2 Αλγόριθμος Διαφοροποίησης Max-sum

Input: Set of candidate results N, size of diverse set k
Output: Set of diverse results $S \subseteq N, |S| = k$
$S = \emptyset$
for $i = 1 \rightarrow \lfloor \frac{k}{2} \rfloor$ **do**
Find $(u, v) = argmax_{x,y \in N}(f_{MAXSUM}(x, y, q))$ ▷ Select pair of docs that maximize Eq 5.10
Set $S = S \cup \{u, v\}$
Set $N = N \setminus \{u, v\}$
end for
if k is odd **then**
$S = S \cup \{i\}, N_i \in N$ ▷ If k is odd add an arbitrary document to S
end if

– **Max-min:** Η μέθοδος Max-min [58] αποσκοπεί στη μεγιστοποίηση της ελάχιστης συνάφειας και ανομοιότητας του επιλεγμένου σετ. Αυτό επιτυγχάνεται με μια άπληστη προσέγγιση, Αλγόριθμος 5.3, που επιλέγει αρχικά ένα ζεύγος εγγράφων που μεγιστοποιούν την εξίσωση 5.11 και, στη συνέχεια, σε κάθε επανάληψη επιλέγει το έγγραφο που μεγιστοποιεί την εξίσωση 5.12:

$$f_{MAXMIN}(u, v, q) = (1 - \lambda)\ (r(u, q) + r(v, q)) + \lambda\ d(u, v) \tag{5.11}$$

$$f_{MAXMIN}(u, q) = \min_{v \in S} d(u, v) \tag{5.12}$$

Ο Αλγόριθμος Max-min 5.3, σε κάθε βήμα, βρίσκει, για κάθε υποψήφιο έγγραφο το πλησιέστερο προς αυτό έγγραφο που ανήκει στο S και υπολογίζει την απόσταση d_{MIN} τους. Το έγγραφο με την μέγιστη απόσταση d_{MIN} εισάγεται στο S.

– **Mono-objective:** Η μέθοδος Mono-objective [58] συνδυάζει τη συνάφεια και τις τιμές ομοιότητας σε μια ενιαία τιμή για κάθε έγγραφο. Ορίζεται ως:

$$f_{MONO}(u, q) = r(u, q) + \frac{\lambda}{|N| - 1} \sum_{v \epsilon N} d(u, v) \tag{5.13}$$

Ο Αλγόριθμος 5.4 περιγράφει την μέθοδο Mono-objective. Ο αλγόριθμος, στο βήμα αρχικοποίησης, υπολογίζει μια βαθμολογία αποστάσεων για κάθε υποψήφιο έγγραφο, με

Αλγόριθμος 5.3 Αλγόριθμος Διαφοροποίησης Max-min

Input: Set of candidate results N, size of diverse set k
Output: Set of diverse results $S \subseteq N, |S| = k$
$S = \emptyset$
Find $(u, v) = argmax_{x,y \in N}(f_{MAXMIN}(x, y, q))$ ▷ initially selects documents that maximize Eq. 5.11
Set $S = S \cup \{u, v\}$
while $|S| < k$ **do**
Find $u = argmax_{x \in N \setminus S}(f_{MAXMIN}(x, q))$ ▷ select document that maximize Eq. 5.12
Set $S = S \cup \{u\}$
end while

βάση την συνάφεια του εκάστοτε εγγράφου με το ερώτημα και την μέση απόσταση του με τα υπόλοιπα έγγραφα. Οι βαθμολογίες αποστάσεων (σκορ) που υπολογίζονται στο στάδιο αρχικοποίησης, δεν ενημερώνονται μετά από κάθε επανάληψη του αλγορίθμου. Έτσι, κάθε βήμα συνίσταται στην επιλογή του εγγράφου με τη μέγιστη βαθμολογία από τα υποψηφία έγγραφα και την τοποθέτησή του στο S.

Αλγόριθμος 5.4 Αλγόριθμος Διαφοροποίησης Mono-objective

Input: Set of candidate results N, size of diverse set k
Output: Set of diverse results $S \subseteq N, |S| = k$
$S = \emptyset$
for $x_i \in N$ **do**
$d(x_i) = f_{MONO}(x, q)$ ▷ Calculate scores based on Eq. 5.13
end for
while $|S| < k$ **do**
Find $u = argmax_{x_i \in N} d(x_i)$ ▷ Sort and select $top - k$ documents
Set $S = S \cup \{u\}$
Set $N = N \setminus \{u\}$
end while

Αλγόριθμοι Περιλήψεων

– **LexRank:** Η μέθοδος LexRank [39] είναι μια στοχαστική μέθοδος, βασισμένη σε γράφο, (stochastic graph-based method) για τον υπολογισμό της σχετικής σημασίας των προτάσεων ενός κειμένου. Η μέθοδος κατατάσσει τις προτάσεις ενός κειμένου σε φθίνουσα σειρά σημασίας, υπό την έννοια της καλύτερης περιγραφής του κειμένου, και έχει προταθεί στην βιβλιογραφία για την αυτόματη κατασκευή περιλήψεων κειμένων.

Βασιζόμενοι στην αναμεταξύ των προτάσεων ενός κειμένου ομοιότητα, ένα έγγραφο δύναται να αναπαρασταθεί ως ένα δίκτυο (γράφος) αλληλένδετων προτάσεων, χρησιμοποιώντας ως τιμές του πίνακα γειτνίασης την ομοιότητα μεταξύ των προτάσεων (Εξίσωση 5.5). Ο πίνακας γειτνίασης ενός πεπερασμένου γράφου G με n κορυφές είναι ο πίνακας διαστάσεων nxn όπου το μη διαγώνιο στοιχείο a_{ij} είναι ο αριθμός ακμών από την κορυ-

φή i στην κορυφή j, και το διαγώνιο στοιχείο a_{ij}, είναι ο αριθμός των ακμών (βρόχοι) από την κορυφή i στον εαυτό της.

Στην περίπτωση που εξετάζουμε, δεν έχουμε προτάσεις ενός εγγράφου, αλλά μια λίστα εγγράφων ταξινομημένη με βάση την συνάφεια με το ερώτημα του χρήστη. Για να μπορέσουμε να χρησιμοποιήσουμε ως μέθοδο διαφοροποίησης την μέθοδο LexRank, χρησιμοποιούμε ως τιμές του πίνακα γειτνίασης την ομοιότητα μεταξύ των εγγράφων στην αρχική λίστα. Στην κατεύθυνση αυτή, θεωρούμε την αρχική λίστα αποτελεσμάτων, ως ένα γράφημα δημιουργώντας συνδέσεις μεταξύ των εγγράφων βασιζόμενοι στο σκορ ομοιότητας τους, όπως στην Εξίσωση 5.5. Χαμηλές τιμές στον πίνακα γειτνίασης, ως αποτέλεσμα ανομοιότητας των εγγράφων, μπορούν να εξαλειφθούν με τον καθορισμό ενός κατωφλίου, έτσι ώστε μόνο σημαντικά παρόμοια έγγραφα να συνδέονται μεταξύ τους. Ωστόσο, όπως σε όλες τις λειτουργίες διακριτοποίησης (discretization), αυτό σημαίνει απώλεια πληροφορίας. Αντί αυτού, επιλέγουμε να αξιοποιήσουμε στο μέγιστο δυνατό βαθμό την ισχύ των συνδέσεων ομοιότητας. Με τον τρόπο αυτό, με βάση την συνάρτηση ομοιότητας των εγγράφων, κατασκευάζουμε ένα δίκτυο ομοιότητας, αποκτώντας ένα πολύ πυκνότερο, αλλά σταθμισμένο γράφημα. Επιπρόσθετα, κανονικοποιούμε τον πίνακα γειτνίασης B, ώστε το άθροισμα της κάθε γραμμής να ισούται με την μονάδα.

Με βάση τα προαναφερθέντα, στην συνάρτηση βαθμολόγησης LexRank 5.14, ο πίνακας B αντικατοπτρίζει την ανά ζεύγος ομοιότητα των εγγράφων και ο τετραγωνικός πίνακας A, που αντιπροσωπεύει την πιθανότητα άλματος σε ένα τυχαία κόμβο στο γράφημα, έχει όλα τα στοιχεία του $1/|N|$, όπου $|N|$ ο αριθμός των εγγράφων.

$$p = [\lambda\ A + (1 - \lambda)\ B]^T p \tag{5.14}$$

Ο Αλγόριθμος LexRank 5.5 εφαρμόζει μια παραλλαγή του PageRank [107] πάνω από ένα δίκτυο εγγράφων. Τυχαίος περίπατος (random walk) στην αλυσίδα Μαρκόφ επιλέγει γειτονική της τρέχουσας, κατάσταση με πιθανότητα $1 - \lambda$ ή άλμα σε οποιαδήποτε κατάσταση στο γράφημα, συμπεριλαμβανομένης της τρέχουσας, με πιθανότητα λ. Η αλυσίδα Μαρκόφ, ή Μαρκοβιανή αλυσίδα, που πήρε το όνομα της από τον Αντρέι Μαρκόφ, είναι ένα μαθηματικό σύστημα που μεταβάλλεται από μια κατάσταση σε μια άλλη, ανάμεσα σε ένα πεπερασμένο αριθμό καταστάσεων.

– **Biased LexRank:** Η μέθοδος Biased LexRank [106] αποτελεί επέκταση της μεθόδου LexRank που λαμβάνει υπόψη τυχόν προηγούμενη κατανομή πιθανοτήτων των εγγράφων π.χ., τη συνάφεια των εγγράφων με ένα συγκεκριμένο ερώτημα. Στην συνάρτηση βαθμολόγησης Biased LexRank, Εξίσωση 5.15, που είναι ανάλογη με αυτήν της LexRank, ο πίνακας A που αντιπροσωπεύει την πιθανότητα άλματος σε ένα τυχαίο κόμβο στο γράφημα, έχει τιμές με βάση την αρχική συνάφεια των εγγράφων με το ερώτημα.

$$p = [\lambda\ A + (1 - \lambda)\ B]^T p \tag{5.15}$$

Αλγόριθμος 5.5 Αλγόριθμος Διαφοροποίησης LexRank

Input: Set of candidate results N, size of diverse set k
Output: Set of diverse results $S \subseteq N, |S| = k$
$S = \emptyset$
$A_{|N||N|} = 1/|N|$
for $u, v \in N$ **do**
$B_{u,v} = d(u, v)$ ▷ Calculate connectivity matrix based on document similarity Eq. 5.5
end for
$p = f_{powermethod}(B, A)$ ▷ Calculate stationary distribution of Eq. 5.14. (Omitted for clarity)
while $|S| < k$ **do**
Find $u = argmax_{x_{i \in N}} p(x_i)$ ▷ Sort and select $top - k$ documents
Set $S = S \cup \{u\}$
Set $N = N \setminus \{u\}$
end while

Ο Αλγόριθμος LexRank 5.5 χρησιμοποιείται και στην περίπτωση της μεθόδου Biased LexRank για την επαναδιατάξη των αποτελεσμάτων αναζήτησης. Ο πίνακας B αντικατοπτρίζει την ανά ζεύγος ομοιότητα των εγγράφων και ο τετραγωνικός πίνακας A, που αντιπροσωπεύει την πιθανότητα άλματος σε ένα τυχαία κόμβο στο γράφημα, έχει τιμές με βάση την αρχική συνάφεια των εγγράφων με το ερώτημα.

Αλγόριθμοι Διαφοροποίησης Γράφων

– **DivRank:** Η μέθοδος DivRank [95] συνδυάζει την δημοτικότητα και την ποικιλομορφία στην κατάταξη των εγγράφων, με βάση ένα χρονικά μεταβαλλόμενο τυχαίο περίπατο (time-variant random walk). Σε αντίθεση με την μέθοδο PageRank [107], η οποία βασίζεται σε σταθερές πιθανότητες, η DivRank υποθέτει ότι πιθανότητες μετάβασης αλλάζουν με την πάροδο του χρόνου και ενισχύονται από τον αριθμό των προηγούμενων επισκέψεων στην ακμή στόχο. Αν $p_T(u, v)$ είναι η πιθανότητα μετάβασης από την ακμή u στην ακμή v κατά το χρόνο T, $p^*(d_j)$ είναι η προηγούμενη κατανομή που καθορίζει την προτίμηση επίσκεψης στην ακμή d_j, και $p_0(u, v)$ είναι η πιθανότητα μετάβασης την ακμή u στην ακμή v πριν από οποιαδήποτε ενίσχυση, τότε:

$$p_T(d_i, d_j) = (1 - \lambda).p^*(d_j) + \lambda.\frac{p_0(d_i, d_j).N_T(d_j)}{D_T(d_i)} \tag{5.16}$$

όπου $N_T(d_j)$ είναι ο αριθμός των επισκέψεων στην ακμή d_j μέχρι την χρονική στιγμή T και

$$D_T(d_i) = \sum_{d_j \in V} p_0(d_i, d_j) N_T(d_j) \tag{5.17}$$

Η μέθοδος DivRank αρχικώς προτάθηκε σε ένα πλαίσιο ανεξάρτητο από το ερώτημα του χρήστη. Ως εκ τούτου, δεν μπορεί να εφαρμοστεί άμεσα στη διαφοροποίηση των

αποτελεσμάτων αναζήτησης. Για τον σκοπό αυτό εισάγουμε μια προ απαίτηση ερωτήματος (query dependent prior) και με τον τρόπο αυτό μπορούμε να χρησιμοποιήσουμε την DivRank στο πλαίσιο της διαφοροποίησης των αποτελεσμάτων αναζήτησης. Στην περίπτωση που εξετάζουμε, χρησιμοποιούμε τα έγγραφα που βρίσκονται στο αρχικό σύνολο ανάκτησης N για ένα δεδομένο ερώτημα q, κατασκευάζουμε το δίκτυο παραπομπών, με βάση τις νομικές αναφορές, μεταξύ των εγγράφων αυτών και εφαρμόζουμε τον Αλγόριθμο DivRank 5.6 για να επιλέξουμε τα $top-k$ διαφοροποιημένα αποτελέσματα στο S.

Αλγόριθμος 5.6 Αλγόριθμος Διαφοροποίησης DivRank

Input: Set of candidate results N, size of diverse set k
Output: Set of diverse results $S \subseteq N, |S| = k$
  $S = \emptyset$
  for $u, v \in N$ **do**
    $B(u,v) = A_{u,v}$ ▷ Connectivity matrix is based on citation network adjacency matrix
  end for
  $p = f_{powermethod}(B)$ ▷ Calculate stationary distribution of Eq. 5.16. (Omitted for clarity)
  while $|S| < k$ **do**
    Find $u = argmax_{x_i \in N} p(x_i)$ ▷ Sort and select $top-k$ documents
    Set $S = S \cup \{u\}$
    Set $N = N \setminus \{u\}$
  end while

– **Grasshopper:** Ένας παρόμοιος με τον DivRank αλγόριθμος κατάταξης περιγράφεται στο [159]. Η μέθοδος Grasshopper ξεκινά με ένα τυχαίο περίπατο, με ομογενή χρόνο, και σε κάθε βήμα, η ακμή με το μεγαλύτερο βάρος ορίζεται σε Κατάσταση Απορρόφησης (absorbing state). Μια κατάσταση i μιας αλυσίδας Μαρκόφ ονομάζεται απορροφητική αν είναι αδύνατο να φύγουμε από αυτή την κατάσταση. Άρα, η κατάσταση i είναι απορροφητική όταν και μόνο όταν $p_{ii} = 1$ και $p_{ij} = 0$ για κάθε $i \neq j$.

$$p_T(d_i, d_j) = (1-\lambda).p^*(d_j) + \lambda.\frac{p_0(d_i, d_j).N_T(d_j)}{D_T(d_i)} \tag{5.18}$$

όπου $N_T(d_j)$ είναι ο αριθμός των επισκέψεων στην ακμή d_j μέχρι την χρονική στιγμή T και

$$D_T(d_i) = \sum_{d_j \in V} p_0(d_i, d_j) N_T(d_j) \tag{5.19}$$

Οι μέθοδοι Grasshopper και DivRank χρησιμοποιούν παρόμοια προσέγγιση και αναμένουμε να παρουσιάσουν παρόμοια/συγκρίσιμα αποτελέσματα. Για τον λόγο αυτό επιλέγουμε να χρησιμοποιήσουμε την μέθοδο Grasshopper με διαφορετικό τρόπο από την μέθοδο DivRank. Αναλυτικότερα, αντί της δημιουργίας του δικτύου παραπομπών μεταξύ των εγγράφων που ανήκουν στο αρχικό σύνολο αποτελεσμάτων, σχηματίζουμε τον πίνακα γειτνίασης με βάση την ομοιότητα των εγγράφου, όπως εξηγήθηκε εκτενώς προηγούμενα, στην περιγραφή της μεθόδου LexRank, Αλγόριθμος 5.5.

5.3 Πειραματική Μελέτη

Στην ενότητα αυτή, περιγράφουμε τη συλλογή νομικών εγγράφων (legal corpus) που χρησιμοποιούμε, τις κατηγορίες ερωτημάτων (query topics) που χρησιμοποιούμε, την μεθοδολογία που σχεδιάσαμε για την αντικειμενική υποσημείωση (annotate) με κρίσεις συνάφειας (relevance judgments) των εγγράφων για κάθε ερώτημα, καθώς και τις μετρικές που χρησιμοποιούμε για την αξιολόγηση της απόδοσης των μεθόδων. Τέλος, παρέχουμε τα αποτελέσματα μαζί με μια σύντομη συζήτηση.

5.3.1 Συλλογή νομικών εγγράφων

Η συλλογή νομικών εγγράφων που χρησιμοποιούμε περιέχει 3890 δικαστικές αποφάσεις από το Ομοσπονδιακό Δικαστήριο της Αυστραλίας - Federal Court of Australia[7]. Οι δικαστικές αυτές αποφάσεις είναι διαθέσιμες από το Αυστραλασιατικό Ινστιτούτο Νομικών Πληροφοριών - Austlii[8] και χρησιμοποιήθηκαν στο [52] για την αξιολόγηση μεθόδων δημιουργίας περιλήψεων δικαστικών αποφάσεων και ανάλυσης νομικών παραπομπών. Η συλλογή περιέχει όλες τις αποφάσεις του Ομοσπονδιακού Δικαστηρίου της Αυστραλίας από το 2006 έως το 2009. Από τις αποφάσεις εξάγαμε το κείμενο και τις νομικές παραπομπές για να τις χρησιμοποιήσουμε στο πλαίσιο διαφοροποίησης που εξετάζουμε.

Το ευρετήριο μας (index) κατασκευάστηκε με τυπική λίστα κοινών λέξεων stop words και τεχνική Porter stemming, με log-based $tf - idf$ σχήμα ευρετηρίασης, έχοντας συνολικά 3890 έγγραφα, 9.782.911 όρους εκ των οποίων 53.791 μοναδικοί όροι.

Ο Πίνακας 5.1 συνοψίζει τις παραμέτρους δοκιμών και τις αντίστοιχες τιμές τους. Κάθε ερώτημα χρήστη υποβάλλεται στο σύστημα και ανακτώνται τα $top - n$ έγγραφα, τα οποία σχηματίζουν το υποψήφιο σύνολο N, με χρήση της Εξίσωσης 5.5 και το log-based $tf - idf$ σχήμα ευρετηρίασης. Η επιλεγμένη τιμή, 100, για το πλήθος του υποψηφίου συνόλου N είναι μια τυπική τιμή που χρησιμοποιείται στη βιβλιογραφία [126]. Οι πειραματικές μελέτες μας πραγματοποιούνται σε μια διπλή στρατηγική: (α) ποιοτική ανάλυση από την άποψη της διαφοροποίησης και της ακρίβειας της κάθε μέθοδος που χρησιμοποιείται σε σχέση με το βέλτιστο σύνολο των αποτελεσμάτων και (β) ανάλυση της επεκτασιμότητας (scalability analysis) των μεθόδων διαφοροποίησης, όταν αυξάνουν οι τιμές των παραμέτρων.

Πίνακας 5.1: Παράμετροι Πειραματικής Μελέτης

Παράμετρος	**Εύρος Τιμών**
Αλγόριθμοι Αξιολόγησης	MMR, Max-min, Max-sum, Mono-objective, LexRank, Biased LexRank, DivRank, Grasshopper
tradeoff λ	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
Αριθμός Εγγράφων $n = \|N\|$	100
Αριθμός Αποτελεσμάτων $k = \|S\|$	5, 10, 20, 30
Υποβληθέντα Ερωτήματα	298

[7] `http://www.fedcourt.gov.au`
[8] `http://www.austlii.edu.au`

5.3.2 Μετρικές Αποτίμησης

Συμφωνώντας με τον ισχυρισμό του [115] ότι 'δεν υπάρχει καθολικά αποδεκτή ως η καλύτερη μετρική αξιολόγησης για την μέτρηση της απόδοσης αλγορίθμων που στοχεύουν στην διαφοροποίηση των αποτελεσμάτων αναζήτησης', επιλέξαμε την αξιολόγηση των μεθόδων διαφοροποίησης χρησιμοποιώντας πλήθος μετρικών που χρησιμοποιούνται στα TREC Diversity Tasks[9]. Συγκεκριμένα, χρησιμοποιούμε:

- a-nDCG: Η μετρική a-normalized discounted cumulative gain [30] ποσοτικοποιεί το ποσοστό μοναδικών πτυχών του ερωτήματος q που καλύπτονται από τα $top-k$ κατεταγμένα έγγραφα. Χρησιμοποιούμε $a = 0.5$, όπως συνήθως χρησιμοποιείται στην αξιολόγηση TREC.

- ERR-IA: Η μετρική Expected reciprocal rank-intent aware [27] βασίζεται στις αλληλοεξαρτήσεις των καταταχθέντων εγγράφων. Η συμβολή του κάθε εγγράφου βασίζεται στην συνάφεια των εγγράφων που κατατάσσονται πριν από αυτό και ως εκ τούτου το τελικό σκορ δεν επηρεάζεται μόνο από την θέση κατάταξης αλλά και από την συνάφεια των προηγουμένως καταταχθέντων εγγράφων.

- S-recall: Η μετρική Subtopic-recal [155] υπολογίζεται ως ο λόγος των μοναδικών πτυχών (aspects) που καλύπτονται από τα $top-k$ αποτελέσματα με τον συνολικό αριθμό των πτυχών. Μετρά σε συγκεκριμένη λίστα αποτελεσμάτων την κάλυψη μιας πτυχής σε βάθος k.

5.3.3 Κρίσεις Συνάφειας

Η πειραματική αξιολόγηση των τεχνικών διαφοροποίησης απαιτεί μια συλλογή εγγράφων, κατηγορίες ερωτημάτων και κρίσεις συνάφειας των εγγράφων για κάθε ερώτημα, οι οποίες ιδανικά έχουν αξιολογηθεί από ειδικούς του χώρου αναφοράς. Μία από τις δυσκολίες στην αξιολόγηση μεθόδων που έχουν σχεδιαστεί για την διαφοροποίηση αποτελεσμάτων αναζήτησης αποτελεί η έλλειψη τυπικών δεδομένων δοκιμών (standard testing data). Παρότι το TREC προσέθεσε εργασία διαφοροποίησης (diversity task) στις εργασίες του διαδικτύου (web track) το 2009, το σύνολο δεδομένων αυτό σχεδιάστηκε υποθέτοντας μια γενική αναζήτηση στο διαδίκτυο, και έτσι, δεν είναι δυνατόν να προσαρμοστεί στις ρυθμίσεις μας. Στην περίπτωση μας, διαθέτοντας μόνο την συλλογή νομικών εγγράφων, χρειάστηκε να καθορίσουμε επιπρόσθετα: (α) τις κατηγορίες ερωτημάτων, (β) μια μέθοδο για τον προσδιορισμό των επιμέρους θεμάτων σε κάθε θέμα και (γ) μια μέθοδο για την υποσημείωση των εγγράφων με κρίσεις συνάφειας για το κάθε θέμα.

Με βάση την έλλειψη τυπικών δοκιμαστικών δεδομένων, ερευνήσαμε για ένα αντικειμενικό τρόπο παραγωγής της λίστας ερωτημάτων και των κρίσεων συνάφειας, για να εκτιμήσουμε και να αξιολογήσουμε τις επιδόσεις των διαφόρων μεθόδων διαφοροποίησης στην συλλογή νομικών εγγράφων. Αναγνωρίζουμε το γεγονός ότι η διαδικασία της αυτόματης παραγωγής

[9] `http://trec.nist.gov/data/web10.html`

ερωτημάτων και κρίσεων συνάφειας αποτελεί, στην καλύτερη των περιπτώσεων, μια ατελή προσέγγιση των ενεργειών ενός πραγματικού χρήστη.

Για το σκοπό αυτό, υλοποιήσαμε την ακόλουθη μέθοδο για την παραγωγή ερωτημάτων και κρίσεων συνάφειας των εγγράφων για κάθε ερώτημα:

Προφίλ χρηστών / ερωτήματα

Χρησιμοποιήσαμε την ταξινόμηση West Law Digest Topics[10] ως ερωτήματα χρήστη. Η West American Digest System είναι μια ταξινόμηση για τον εντοπισμό και την οργάνωση με βάση την θεματική περιοχή δικαστικών αποφάσεων. Χρησιμοποιείται για την οργάνωση ολόκληρου του σώματος του Αμερικανικού δικαίου.

Αναλυτικότερα, κάθε θέμα της ταξινόμησης, υποβλήθηκε ως ερώτημα χρήστη στο σύστημα ανάκτησης εγγράφων. Τα ερωτήματα ακραίων τιμών, είτε πολύ γενικά ή πολύ συγκεκριμένα/σπάνια, αφαιρέθηκαν χρησιμοποιώντας το διατεταρτημοριακό διάστημα (interquartile range). Δηλαδή αφαιρέσαμε τιμές μεγαλύτερες από $Q1$ και μικρότερες από $Q3$, ακολουθιακά σε όρους πλήθους συναφών εγγράφων και κατανομής συνάφειας στα αποτελέσματα της αναζήτησης, απαιτώντας παράλληλα ελάχιστη κάλυψη $min|N|$ αποτελεσμάτων. Συνολικά, διατηρήσαμε 289 ερωτήματα. Ο Πίνακας 5.2 παρέχει ένα δείγμα από τα θέματα που θεωρούμε περαιτέρω ως ερωτήματα των χρηστών.

Πίνακας 5.2: Δείγμα ερωτημάτων με βάση την ταξινόμηση West Law Digest Topics

1:	Abandoned and Lost Property	3:	Abortion and Birth Control
24:	Aliens Immigration and Citizenship	31:	Antitrust and Trade Regulation
61:	Breach of Marriage Promise	84:	Commodity Futures Trading Regulation
88:	Compromise and Settlement	199:	Implied and Constructive Contracts
291:	Privileged Communications and Confidentiality	363:	Threats Stalking and Harassment

Επισημειώσεις ερωτημάτων και Σύνολο Αντικειμενικής Αλήθειας

Για κάθε ένα από τα ερωτήματα, κρατήσαμε τα $top-n$ αποτελέσματα. Εν συνεχεία εκπαιδεύσαμε, στα $top-n$ αποτελέσματα για κάθε ερώτηση, ένα μοντέλο με βάση τον αλγόριθμο λανθάνουσας κατανομής Dirichlet - LDA [19] (LDA topic model), χρησιμοποιώντας την εφαρμογή ανοικτού λογισμικού Mallet[11]. Η μοντελοποίηση αυτή μας παρέχει έναν τρόπο να συναχθεί η λανθάνουσα δομή πίσω από μια συλλογή εγγράφων αφού κάθε παραγόμενο θέμα αποτελείται από ένα σύνολο από λέξεις κλειδιά με αντίστοιχα βάρη. Ο Πίνακας 5.3 παρέχει ένα δείγμα από τις κορυφαίες λέξεις-κλειδιά για κάθε θέμα με βάση τα αποτελέσματα για το Ερώτημα 1: Abandoned and Lost Property.

[10] https://en.wikipedia.org/wiki/West_American_Digest_System
[11] `http://mallet.cs.umass.edu/`

Πίνακας 5.3: Δείγμα από τις κορυφαίες λέξεις-κλειδιά για κάθε θέμα με βάση τα αποτελέσματα για το Ερώτημα 1: Abandoned and Lost Property

Θέμα	**Κορυφαίες Λέξεις**
1	court applicant property respondent claim order costs company trustee trust
2	evidence agreement contract business dr time hamilton respondent applicant sales
3	tribunal rights land title native evidence area claim interests appellant
4	evidence agreement meeting conduct time tiltform second brand september australia
5	evidence professor university dr property gray patent uwaclaim nrc

Με βάση την προκύπτουσα κατανομή των θεμάτων, με όριο αποδοχής το 20%, επισημαίνουμε ένα έγγραφο ως σχετικό για το εν λόγο θέμα. Η τιμή του ορίου αποδοχής 20% στην κατανομή των θεμάτων από την μέθοδο LDA για την επισήμανση των εγγράφων με τα λανθάνοντα θέματα, επηρεάζει την πυκνότητα κατανομής θεμάτων και όχι την γενίκευση της χρησιμοποιούμενης μεθόδου ή το αποτέλεσμα της αξιολόγησης. Πραγματοποιήσαμε πειράματα με άλλες ρυθμίσεις του ορίου αποδοχής, 10% και 30% αντίστοιχα, με συγκρίσιμα αποτελέσματα.

Ο Πίνακας 5.4 παρέχει ένα δείγμα της πιθανοτικής κατανομής θεμάτων για πέντε τυχαία έγγραφα από το σύνολο των αποτελεσμάτων για το Ερώτημα 1: Abandoned and Lost Property. Με βάση τα στοιχεία του πίνακα, με όριο αποδοχής το 20%, μπορούμε να συμπεράνουμε ότι το έγγραφο με αριθμό 08_711 καλύπτει τα θέματα 1 και 2, ενώ το έγγραφο με αριθμό 07_924 καλύπτει τα θέματα 1 και 5.

Πίνακας 5.4: Πιθανοτική κατανομή θεμάτων για πέντε τυχαία έγγραφα από το σύνολο των αποτελεσμάτων για το Ερώτημα 1: Abandoned and Lost Property.

Έγγραφο	**Θέμα 1**	**Θέμα 2**	**Θέμα 3**	**Θέμα 4**	**Θέμα 5**
08_711	0.731	0.267	0.00172	8.79×10^{-5}	4×10^{-5}
09_1395	0.467	2.25×10^{-5}	0.459	0.0746	1.04×10^{-5}
09_1383	0.99	2.85×10^{-5}	0.00944	0.000298	1.31×10^{-5}
06_1169	0.994	5.33×10^{-5}	0.00559	5.41×10^{-5}	2.46×10^{-5}
07_924	0.237	4.83×10^{-5}	4.64×10^{-5}	4.9×10^{-5}	0.763

Συνεπώς, από την αντίστοιχη θεματική κατανομή, αποκτούμε δυαδικές εκτιμήσεις για κάθε ερώτημα και έγγραφο στο σύνολο των αποτελεσμάτων. Με τον τρόπο αυτό, χρησιμοποιώντας την μέθοδο LDA, δημιουργήσαμε ένα Σύνολο Αντικειμενικής Αλήθειας (Ground truth), αποτελούμενο από δυαδικές υποσημειώσεις για κάθε ερώτημα. Με γνώμονα την προώθηση σχετικών θεμάτων και την ενίσχυση της συνεργασίας σε θέματα διαφοροποιημένης ανάκτησης νομικών κειμένων, έχουμε διαθέσει ελεύθερα στο διαδίκτυο [12] το πλήρες σύνολο δεδομένων, τα δεδομένα Ground truth, τα ερωτήματα και τις επισημειώσεις των εγγράφων για κάθε ερώτημα.

[12] https://github.com/mkoniari/LegalDivEval

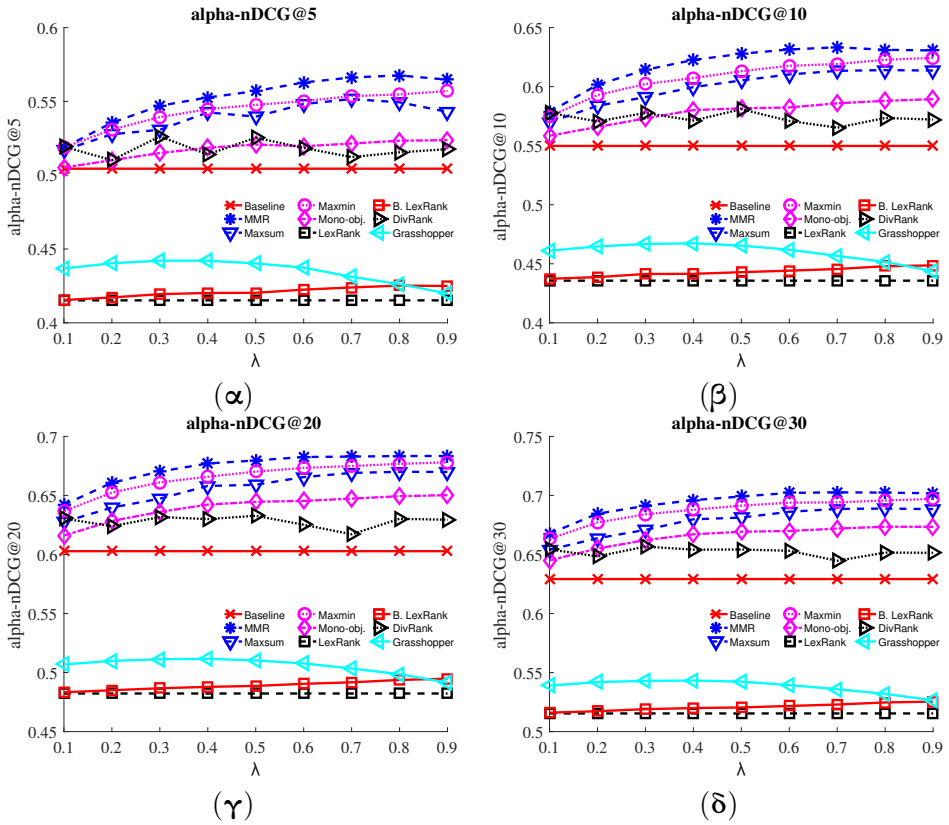
5.3.4 Αποτελέσματα

Ως βάση (baseline) για την σύγκριση των μεθόδων διαφοροποίησης, θεωρούμε την αρχική κατάταξη των εγγράφων, που παράγεται από την συνάρτηση ομοιότητας εγγράφου με το ερώτημα του χρήστη (Εξίσωση 5.8) και το log-based $tf - idf$ σχήμα ευρετηρίασης. Για κάθε ερώτημα, το αρχικό σύνολο N περιέχει τα $top - n$ αποτελέσματα. Η παράμετρος trade-off $\lambda \in [0..1]$ ρυθμίζεται με βήμα 0.1 για κάθε μέθοδο ξεχωριστά. Παρουσιάζουμε τα αποτελέσματα της αξιολόγησης για τις μεθόδους που αξιολογούμε, χρησιμοποιώντας τις προαναφερθείσες μετρικές αξιολόγησης, με τιμές cut-off @5, @10, @20, @30, όπως είθισται στις αξιολογήσεις TREC. Τα αποτελέσματα παρουσιάζονται με σταθερή παράμετρο $n = |N|$. Σημειώνουμε ότι κάθε μία από τις δοκιμές διαφοροποίησης εφαρμόζεται σε συνδυασμό με κάθε ένα από τους αλγορίθμους διαφοροποίησης και για κάθε ερώτημα του χρήστη.

Το Σχήμα 5.2 απεικονίζει τις τιμές a-normalized discounted cumulative gain (a-nDCG) της κάθε μεθόδου για διαφορετικές τιμές του λ. Οι αλγόριθμοι διαφοροποίησης αποτελεσμάτων αναζήτησης (MMR, Max-sum, Max-min και Mono-objective) βελτιώνουν την μέτρηση βάσης (baseline), ενώ αντίθετα, οι μέθοδοι περιλήψεων (LexRank, Biased LexRank) και η μέθοδος Grasshopper, όπως χρησιμοποιήθηκε χωρίς την χρήση του δικτύου νομικών παραπομπών, αποτυγχάνουν να βελτιώσουν την μέτρηση βάσης, επιφέροντας χειρότερα αποτελέσματα από την αρχική τιμή κατάταξης σε όλα τα επίπεδα για όλες τις μετρήσεις. Τα αποτελέσματα του αλγόριθμου διαφοροποίησης γράφων DivRank ποικίλλουν σημαντικά ανάλογα με τις διάφορες τιμές του λ. Αποδίδουμε το εύρημα αυτό στο ακραία (extreme) αραιό δίκτυο νομικών παραπομπών, καθώς η συλλογή νομικών εγγράφων που χρησιμοποιούμε καλύπτει σύντομο χρονικό διάστημα (τρία χρόνια).

Το Σχήμα 5.3 απεικονίζει τις τιμές normalized expected reciprocal rank-intent aware (nERR-IA) της κάθε μεθόδου για διαφορετικές τιμές του λ. Οι αλγόριθμοι διαφοροποίησης αποτελεσμάτων αναζήτησης (MMR, Max-sum, Max-min και Mono-objective) επιφέρουν βελτίωση των αποτελεσμάτων της μέτρησης βάσης. Ειδικότερα καθώς αυξάνουν οι τιμές του λ, αυξάνει η προτίμηση για διαφοροποίηση των αποτελεσμάτων και οι τιμές nERR-IA για όλες τις μεθόδους. Οι μέθοδοι περιλήψεων (LexRank, Biased LexRank) και η μέθοδος Grasshopper αποτυγχάνουν και σε αυτήν την περίπτωση να βελτιώσουν την μέτρηση βάσης, επιφέροντας χειρότερα αποτελέσματα από την αρχική τιμή κατάταξης σε όλα τα επίπεδα για όλες τις μετρήσεις. Τα αποτελέσματα του αλγόριθμου διαφοροποίησης γράφων DivRank, όπως και στην προηγούμενη περίπτωση, ποικίλλουν σημαντικά ανάλογα με τις διάφορες τιμές του λ. Η μέθοδος MMR επιτυγχάνει τα καλύτερα αποτελέσματα, σε σχέση με τις υπόλοιπες μεθόδους. Παρατηρούμε επίσης, ότι η μέθοδος Max-min αποδίδει καλύτερα από την Max-sum. Υπάρχουν λίγες περιπτώσεις όπου και οι δύο μέθοδοι παρουσιάζουν σχεδόν παρόμοια απόδοση, ειδικά σε χαμηλότερα επίπεδα ανάκλησης (π.χ., για nERR-IA @5 και τιμές λ 0.1, 0.4, 0.6, 0.7). Επίσης και σε αυτήν την περίπτωση η μέθοδος Mono-objective παρουσιάζει χαμηλότερη απόδοση σε σύγκριση με τις μεθόδους MMR, Max-sum, Max-min στην μετρική nERR-IA για όλες τις τιμές λ.

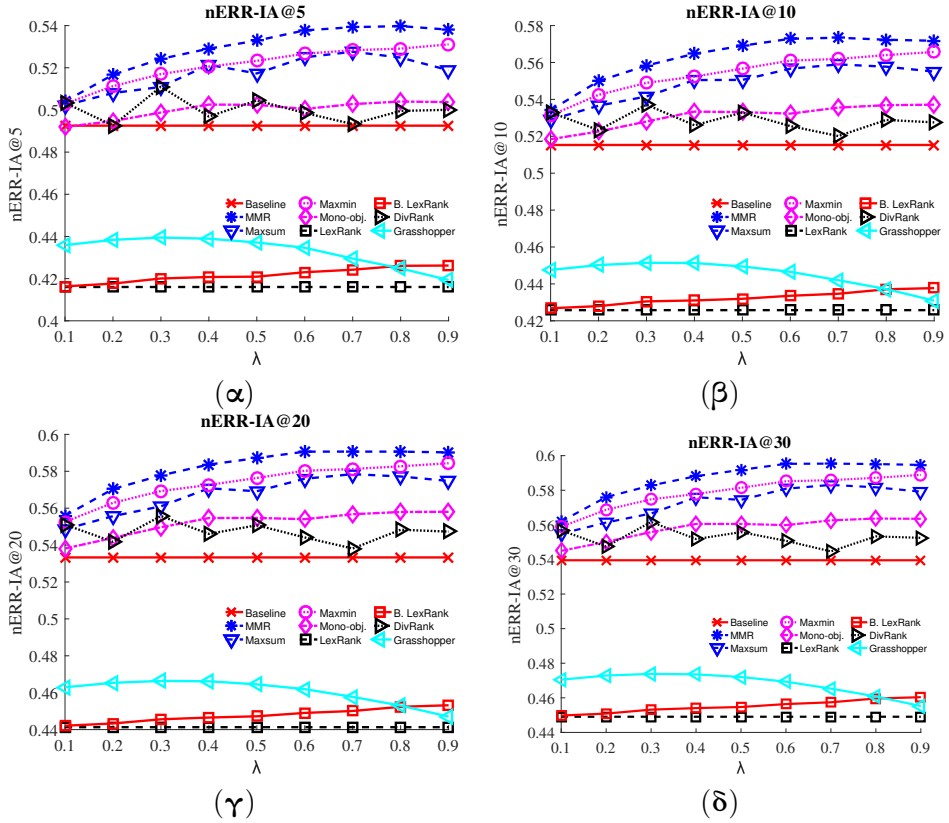
Το Σχήμα 5.4 απεικονίζει τις τιμές subtopic-recall (S-recall), σε διάφορα επίπεδα @5,



Σχήμα 5.2: Τιμές *a*-Normalized discounted cumulative gain (a-nDCG) σε διάφορα επίπεδα @5, @10, @20, @30 για μέτρηση βάσης και μεθόδους MMR, Max-sum, Max-min, Mono-objective, LexRank, Biased LexRank, DivRank και Grasshopper. (α) alpha-nDCG@5 (β) alpha-nDCG@10 (γ) alpha-nDCG@20 (δ) alpha-nDCG@30. (Καλύτερη απεικόνιση έγχρωμα.)

@10, @20, @30, της κάθε μεθόδου για διαφορετικές τιμές του λ. Παρατηρούμε ότι και σε αυτήν περίπτωση, οι αλγόριθμοι διαφοροποίησης αποτελεσμάτων αναζήτησης (MMR, Max-sum, Max-min και Mono-objective) επιφέρουν βελτίωση των αποτελεσμάτων της μέτρησης βάσης. Ειδικότερα καθώς αυξάνουν οι τιμές του λ, αυξάνει η προτίμηση για διαφοροποίηση των αποτελεσμάτων και οι τιμές nERR-IA για όλες τις μεθόδους. Στα χαμηλότερα επίπεδα (π.χ., @5, @10), η μέθοδος MMR υπερτερεί των υπόλοιπων μεθόδων, ενώ για τα ανώτερα επίπεδα (π.χ., @20, @30), οι μέθοδοι MMR και Max-min έχουν συγκρίσιμα αποτελέσματα. Παρατηρούμε επίσης ότι η μέθοδος Max-min αποδίδει καλύτερα, σε σχέση με την μέθοδο Max-sum, η οποία επιτυγχάνει συνεχώς καλύτερα αποτελέσματα από ό, τι η μέθοδος Mono-objective. Τέλος οι μέθοδοι περιλήψεων (LexRank, Biased LexRank) και η μέθοδος Grasshopper αποτυγχάνουν και σε αυτήν την περίπτωση να βελτιώσουν την μέτρηση βάσης, επιφέροντας χειρότερα αποτελέσματα από την αρχική τιμή κατάταξης σε όλα τα επίπεδα για όλες τις μετρήσεις. Συνολικά, παρατηρούμε παρόμοια συμπεριφορά των μεθόδων με αυτή που αναλύθηκε στα Σχήματα 5.2 και 5.3.

Συμπερασματικά, μεταξύ όλων των μετρικών, διαπιστώνουμε παρόμοια συμπεριφορά στις τάσεις των γραφικών παραστάσεων. Οι MMR, Max-sum, Max-min, Mono-objective και Di-

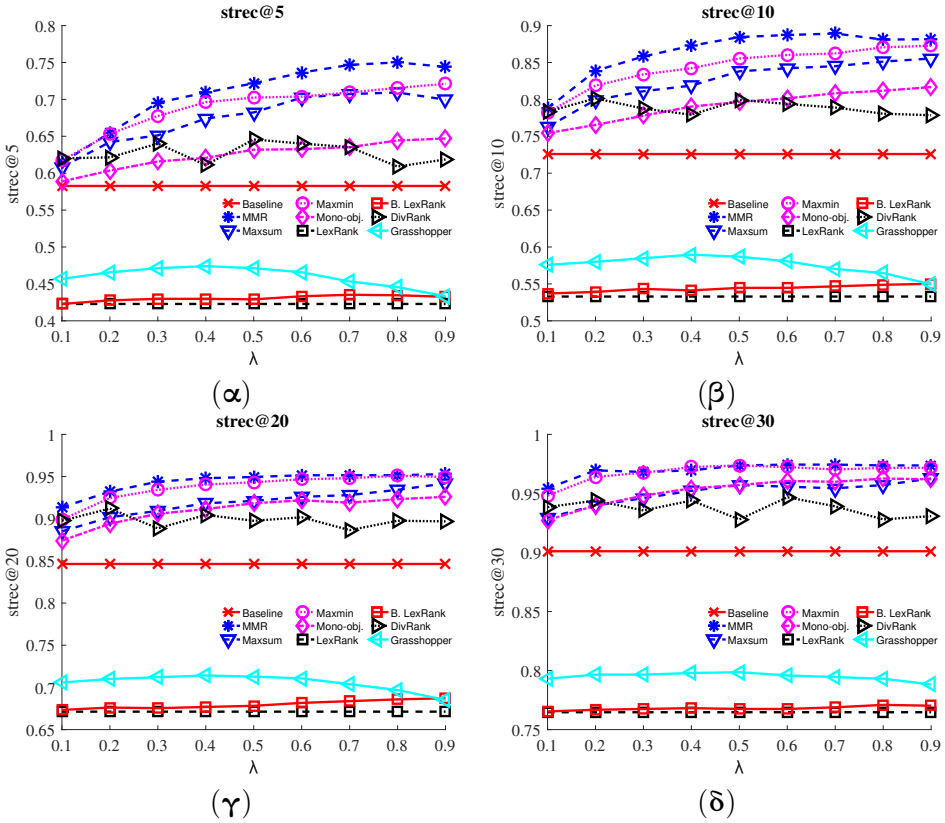


Σχήμα 5.3: Τιμές Normalized expected reciprocal rank-intent aware (nERR-IA) σε διάφορα επίπεδα @5, @10, @20, @30 για μέτρηση βάσης και μεθόδους MMR, Max-sum, Max-min, Mono-objective, LexRank, Biased LexRank, DivRank και Grasshopper. (α) nERR-IA@5 (β) nERR-IA@10 (γ) nERR-IA@20 (δ) nERR-IA@30. (Καλύτερη απεικόνιση έγχρωμα.)

vRank επιφέρουν στατιστικά σημαντικές, χρησιμοποιώντας τον έλεγχο t κατά ζεύγη (paired two-sided t-test), βελτιώσεις όλων των μετρικών, παρέχοντας στους ενδιαφερόμενους φορείς νομικής πληροφορίας μια ευρύτερη εικόνα του νομικού χώρου των αποτελεσμάτων. Επιπρόσθετα οι τάσεις των γραφικών παραστάσεων - μετρικών, τονίζουν τα όρια ισορροπίας για τα νομικά συστήματα μεταξύ της συνάφειας των αποτελεσμάτων αναζήτησης ή την δειγματοληψία του χώρου πληροφορίες γύρω από το νομικό ερώτημα.

Ο Πίνακας Α΄.1 συνοψίζει το μέσο όρο (average) αποτελεσμάτων των υπό αξιολόγηση μεθόδων διαφοροποίησης. Στατιστικώς σημαντικές τιμές, χρησιμοποιώντας τον έλεγχο t κατά ζεύγη (paired two-sided t-test), με τιμές $p_{value} < 0.05$ σημειώνονται με $^\circ$ και για $p_{value} < 0.01$ με *.

Η αποτελεσματικότητα των μεθόδων διαφοροποίησης επίσης απεικονίζεται στον Πίνακα 5.5, στον οποίο παρουσιάζονται τα κορυφαία πέντε (5) αποτελέσματα για τρία (3) τυχαία ερωτήματα, χρησιμοποιώντας την συλλογή νομικών εγγράφων μας ($|S| = 30$ ανδ $|N| = 100$) για $\lambda = 0$ (όχι διαφοροποίηση), $\lambda = 0.1$ (ελαφριά διαφοροποίηση), $\lambda = 0.5$ (μέτρια διαφοροποίηση) και $\lambda = 0.9$ (υψηλή διαφοροποίηση). Παρουσιάζουμε μόνο τα αποτελέσματα της μεθόδου MMR καθώς, σχεδόν σε όλες τις παραλλαγές, υπερτερεί των υπόλοιπων προ-



Σχήμα 5.4: Τιμές Subtopic recall (S-recall) σε διάφορα επίπεδα @5, @10, @20, @30 για μέτρηση βάσης και μεθόδους MMR, Max-sum, Max-min, Mono-objective, LexRank, Biased LexRank, DivRank και Grasshopper. (α) S-recall@5 (β) S-recall@10 (γ) S-recall20 (δ) S-recall@30. (Καλύτερη απεικόνιση έγχρωμα.)

σεγγίσεων. Λόγω περιορισμών χώρου, παραθέτουμε μόνο τον τίτλο της δικαστικής απόφασης για κάθε περίπτωση/εγγραφή. Η υπερ-σύνδεση για το πλήρες κείμενο κάθε καταχώρισης μπορεί να σχηματιστεί χρησιμοποιώντας τη μορφή `http://www.austlii.edu.au/au/cases/cth/FCA/{year}/{docNum}.html` και αντικαθιστώντας τις τιμές που δίνονται στον Πίνακα 5.5. Για παράδειγμα, για την πρώτη εγγραφή του Πίνακα 5.5 ο αντίστοιχος υπερσύνδεσμος, αντικαθιστώντας τις τιμές 2008 και 1503 για τα πεδία *year* και *docNum* είναι `http://www.austlii.edu.au/au/cases/cth/FCA/2008/1503.html`

Για τιμές $\lambda = 0$, ο πίνακας περιέχει τα κορυφαία 5 αποτελέσματα του S διατεταγμένα με βάση την συνάφεια με το ερώτημα. Τα αποτελέσματα αυτά περιέχουν πολλές παρόμοιες, σχεδόν διπλότυπες δικαστικές αποφάσεις, με βάση τον τίτλο της κάθε απόφασης. Καθώς οι τιμές του λ αυξάνουν, λιγότερα διπλότυπα απαντώνται στην λίστα των κορυφαίων 5 αποτελεσμάτων και τα αποτελέσματα καλύπτουν περισσότερα θέματα, με βάση τον τίτλο της κάθε απόφασης. Παρατηρούμε επίσης ότι η λίστα των αποτελεσμάτων με υψηλή διαφοροποίηση περιέχει τίτλους δικαστικών αποφάσεων που περιέχουν σχεδόν όλους τους όρους του έκαστου ερωτήματος, γεγονός που φανερώνει ότι η δικαστική απόφαση έχει σχέση με διάφορα θέματα μεταξύ των άλλων περιπτώσεων στο σύνολο των αποτελεσμάτων.

Πίνακας 5.5: Κορυφαία πέντε αποτελέσματα για τρία τυχαία ερωτήματα, με την μέθοδο MMR, χρησιμοποιώντας την συλλογή νομικών εγγράφων μας ($|S| = 30$ και $|N| = 100$) για $\lambda = 0$ (όχι διαφοροποίηση), $\lambda = 0.1$ (ελαφριά διαφοροποίηση), $\lambda = 0.5$ (μέτρια διαφοροποίηση) και $\lambda = 0.9$ (υψηλή διαφοροποίηση).

α/α	**Συνάφεια** ($\lambda = 0$)	**MMR**: ελαφριά διαφοροποίηση ($\lambda = .1$)	**MMR**: μεσαία διαφοροποίηση ($\lambda = .5$)	**MMR**: υψηλή διαφοροποίηση ($\lambda = .9$)
Ερώτημα 24: Aliens Immigration and Citizenship				
1	Virgin Holdings SA v Commissioner of Taxation [2008] FCA 1503 (10 October 2008)	Virgin Holdings SA v Commissioner of Taxation [2008] FCA 1503 (10 October 2008)	Virgin Holdings SA v Commissioner of Taxation [2008] FCA 1503 (10 October 2008)	Virgin Holdings SA v Commissioner of Taxation [2008] FCA 1503 (10 October 2008)
2	Undershaft (No 1) Limited v Commissioner of Taxation [2009] FCA 41 (3 February 2009)	Fowler v Commissioner of Taxation [2008] FCA 528 (21 April 2008)	Fowler v Commissioner of Taxation [2008] FCA 528 (21 April 2008)	Soh v Commonwealth of Australia [2008] FCA 520 (18 April 2008)
3	Fowler v Commissioner of Taxation [2008] FCA 528 (21 April 2008)	Wight v Honourable Chris Pearce, MP, Parliamentary Secretary to the Treasurer [2007] FCA 26 (29 January 2007)	Coleman v Minister for Immigration & Citizenship [2007] FCA 1500 (27 September 2007)	SZJDI v Minister for Immigration & Citizenship (No. 2) [2008] FCA 813 (16 May 2008)
4	Wight v Honourable Chris Pearce, MP, Parliamentary Secretary to the Treasurer [2007] FCA 26 (29 January 2007)	Undershaft (No 1) Limited v Commissioner of Taxation [2009] FCA 41 (3 February 2009)	Charlie v Minister for Immigration and Citizenship [2008] FCA 1025 (10 July 2008)	Charlie v Minister for Immigration and Citizenship [2008] FCA 1025 (10 July 2008)
5	Coleman v Minister for Immigration & Citizenship [2007] FCA 1500 (27 September 2007)	Coleman v Minister for Immigration & Citizenship [2007] FCA 1500 (27 September 2007)	VSAB v Minister for Immigration and Multicultural and Indigenous Affairs [2006] FCA 239 (17 March 2006)	VSAB v Minister for Immigration and Multicultural and Indigenous Affairs [2006] FCA 239 (17 March 2006)
Ερώτημα 84: Commodity Futures Trading Regulation				
1	BHP Billiton Iron Ore Pty Ltd v The National Competition Council [2006] FCA 1764 (18 December 2006)	BHP Billiton Iron Ore Pty Ltd v The National Competition Council [2006] FCA 1764 (18 December 2006)	BHP Billiton Iron Ore Pty Ltd v The National Competition Council [2006] FCA 1764 (18 December 2006)	BHP Billiton Iron Ore Pty Ltd v The National Competition Council [2006] FCA 1764 (18 December 2006)
2	Australian Securities & Investments Commission v Lee [2007] FCA 918 (15 June 2007)	Australian Securities & Investments Commission v Lee [2007] FCA 918 (15 June 2007)	Australian Securities & Investments Commission v Lee [2007] FCA 918 (15 June 2007)	Australian Competition and Consumer Commission v Dally M Publishing and Research Pty Limited [2007] FCA 1220 (10 August 2007)

Συνέχεια στην επόμενη σελίδα ...

Πίνακας 5.5 – *Συνέχεια από την προηγούμενη σελίδα*

	Συνάφεια $\lambda = 0$	**MMR**: ελαφριά διαφοροποίηση ($\lambda = .1$)	**MMR**: μεσαία διαφοροποίηση ($\lambda = .5$)	**MMR**: υψηλή διαφοροποίηση ($\lambda = .9$)
3	Woodside Energy Ltd (ABN 63 005 482 986) v Commissioner of Taxation (No 2) [2007] FCA 1961 (10 December 2007)	Woodside Energy Ltd (ABN 63 005 482 986) v Commissioner of Taxation (No 2) [2007] FCA 1961 (10 December 2007)	Woodside Energy Ltd (ABN 63 005 482 986) v Commissioner of Taxation (No 2) [2007] FCA 1961 (10 December 2007)	Heritage Clothing Pty Ltd trading as Peter Jackson Australia v Mens Suit Warehouse Direct Pty Ltd trading as Walter Withers [2008] FCA 1775 (28 November 2008)
4	BHP Billiton Iron Ore Pty Ltd v National Competition Council (No 2) [2007] FCA 557 (19 April 2007)	Keynes v Rural Directions Pty Ltd (No 2) (includes Corrigendum dated 16 July 2009) [2009] FCA 567 (3 June 2009)	Keynes v Rural Directions Pty Ltd (No 2) (includes Corrigendum dated 16 July 2009) [2009] FCA 567 (3 June 2009)	Travelex Limited v Commissioner of Taxation (Corrigendum dated 4 February 2009) [2008] FCA 1961 (19 December 2008)
5	Keynes v Rural Directions Pty Ltd (No 2) (includes Corrigendum dated 16 July 2009) [2009] FCA 567 (3 June 2009)	Queanbeyan City Council v ACTEW Corporation Limited [2009] FCA 943 (24 August 2009)	Heritage Clothing Pty Ltd trading as Peter Jackson Australia v Mens Suit Warehouse Direct Pty Ltd trading as Walter Withers [2008] FCA 1775 (28 November 2008)	Ashwick (Qld) No 127 Pty Ltd (ACN 010 577 456) v Commissioner of Taxation [2009] FCA 1388 (26 November 2009)
Ερώτημα 291: Privileged Communications and Confidentiality				
1	Siam Polyethylene Co Ltd v Minister of State for Home Affairs (No 3) [2009] FCA 839 (7 August 2009)	Siam Polyethylene Co Ltd v Minister of State for Home Affairs (No 3) [2009] FCA 839 (7 August 2009)	Siam Polyethylene Co Ltd v Minister of State for Home Affairs (No 3) [2009] FCA 839 (7 August 2009)	Siam Polyethylene Co Ltd v Minister of State for Home Affairs (No 3) [2009] FCA 839 (7 August 2009)
2	AWB Limited v Australian Securities and Investments Commission [2008] FCA 1877 (11 December 2008)	AWB Limited v Australian Securities and Investments Commission [2008] FCA 1877 (11 December 2008)	AWB Limited v Australian Securities and Investments Commission [2008] FCA 1877 (11 December 2008)	Krueger Transport Equipment Pty Ltd v Glen Cameron Storage [2008] FCA 803 (30 May 2008)
3	Brookfield Multiplex Limited v International Litigation Funding Partners Pte Ltd (No 2) [2009] FCA 449 (6 May 2009)	Brookfield Multiplex Limited v International Litigation Funding Partners Pte Ltd (No 2) [2009] FCA 449 (6 May 2009)	Autodata Limited v Boyce's Automotive Data Pty Limited [2007] FCA 1517 (4 October 2007)	Futuretronics.com.au Pty Limited v Graphix Labels Pty Ltd [2007] FCA 1621 (29 October 2007)
4	Cadbury Schweppes Pty Ltd (ACN 004 551 473) v Amcor Limited (ACN 000 017 372) [2008] FCA 88 (19 February 2008)	Barrett Property Group Pty Ltd v Carlisle Homes Pty Ltd (No 2) [2008] FCA 930 (17 June 2008)	Barrett Property Group Pty Ltd v Carlisle Homes Pty Ltd (No 2) [2008] FCA 930 (17 June 2008)	Australian Competition & Consumer Commission v Visy Industries [2006] FCA 136 (23 February 2006)

Συνέχεια στην επόμενη σελίδα ...

Πίνακας 5.5 – *Συνέχεια από την προηγούμενη σελίδα*

	Συνάφεια $\lambda = 0$	**MMR**: ελαφριά διαφοροποίηση ($\lambda = .1$)	**MMR**: μεσαία διαφοροποίηση ($\lambda = .5$)	**MMR**: υψηλή διαφοροποίηση ($\lambda = .9$)
5	Barrett Property Group Pty Ltd v Carlisle Homes Pty Ltd (No 2) [2008] FCA 930 (17 June 2008)	Cadbury Schweppes Pty Ltd (ACN 004 551 473) v Amcor Limited (ACN 000 017 372) [2008] FCA 88 (19 February 2008)	Optus Networks Ltd v Telstra Corporation Ltd (No. 2) (includes Corrigendum dated 7 July 2009) [2009] FCA 422 (9 July 2009)	IO Group Inc v Prestige Club Australasia Pty Ltd (No 2) [2008] FCA 1237 (11 August 2008)

5.4 Πειραματική Μελέτη με Κριτήρια Διαφοροποίησης

Σκοπός μας είναι η πειραματική αξιολόγηση των τεχνικών διαφοροποίησης που μελετάμε τόσο σε διαφορετική κατηγορία τύπου νομικού συστήματος, χρησιμοποιώντας διαφορετική συλλογή νομικών εγγράφων, όσο και διαφορετικά σενάρια διαφοροποίησης, αξιολογώντας ταυτόχρονα την συνεισφορά των κριτηρίων διαφοροποίησης που προτείνουμε.

Στην ενότητα αυτή, περιγράφουμε τη συλλογή νομικών εγγράφων που χρησιμοποιούμε, τις κατηγορίες ερωτημάτων που χρησιμοποιούμε, την μεθοδολογία που σχεδιάσαμε για την αντικειμενική υποσημείωση με κρίσεις συνάφειας των εγγράφων για κάθε ερώτημα, καθώς και τις μετρικές που χρησιμοποιούμε για την αξιολόγηση της απόδοσης των μεθόδων. Τέλος, παρέχουμε τα αποτελέσματα μαζί με μια σύντομη συζήτηση.

5.4.1 Συλλογή νομικών εγγράφων

Η συλλογή νομικών εγγράφων που χρησιμοποιούμε περιέχει 63,742 δικαστικές αποφάσεις από το Ανώτατο Δικαστήριο των Ηνωμένων Πολιτειών Αμερικής (Η.Π.Α.) - Supreme Court of the United States[13]. Οι δικαστικές αυτές αποφάσεις είναι διαθέσιμες από το CourtListener[14], υπηρεσία που παρέχει δωρεάν πρόσβαση σε αποφάσεις και γνωμοδοτήσεις ομοσπονδιακών και πολιτειακών δικαστηρίων των Η.Π.Α.. Η συλλογή νομικών εγγράφων που χρησιμοποιούμε περιέχει όλες τις αποφάσεις του Ανωτάτου Δικαστηρίου των Η.Π.Α., από το 1754 μέχρι το 2015, καλύπτοντας περισσότερο από δύο αιώνες νομικής ιστορίας. Από τις αποφάσεις εξάγαμε το κείμενο καθώς και όλες τις απαραίτητες πληροφορίες για την εφαρμογή των κριτηρίων διαφοροποίησης που προτείνουμε, όπως σχέσεις με άλλα έγγραφα (νομικές παραπομπές), ημερομηνία εκδόσεως της αποφάσεως.

Καθώς η συλλογή νομικών εγγράφων ήταν αρχικά αταξινόμητη, προκειμένου να επισημειώσουμε τις δικαστικές αποφάσεις με θεματικές κατηγορίες, χρησιμοποιήσαμε την Βάση Δεδομένων του Ανωτάτου Δικαστηρίου - Supreme Court Database[15], ενώνοντας τις αντίστοιχες εγγραφές με βάση την κοινά χρησιμοποιούμενη παράμετρο - κωδικός απόφασης SCDB Case ID. Επισημαίνετε ότι οι θεματικές κατηγοριοποιήσεις της Βάσης Δεδομένων του Ανωτάτου

[13] `http://www.supremecourt.gov/`
[14] `http://www.courtlistener.com`
[15] `http://scdb.wustl.edu`

Δικαστηρίου αποτελούν προϊόν εργασίας ομάδας εμπειρογνωμόνων, οι οποίοι αναλύουν και ερμηνεύουν τις νομικές διατάξεις της κάθε απόφασης.

Η προ-επεξεργασία της συλλογής νομικών εγγράφων περιλάμβανε επίσης την αφαίρεση κοινών λέξεων και τεχνική Porter stemming. Το ευρετήριο που σχηματίστηκε, με log-based $tf - idf$ σχήμα ευρετηρίασης, περιλαμβάνει τελικά 63,742 έγγραφα, 54,243,977 όρους εκ των οποίων 174,370 είναι μοναδικοί όροι. Συνολικά, θεωρούμε ότι η συλλογή νομικών εγγράφων που χρησιμοποιούμε, η οποία καλύπτει δυόμιση αιώνες αποφάσεων ανώτατου δικαστηρίου, έχει κατάλληλο μέγεθος για την αξιολόγηση της αποτελεσματικότητας της προτεινόμενης προσέγγισης.

5.4.2 Μετρικές Αποτίμησης

Όπως και στην προαναφερθείσα πειραματική μελέτη (Ενότητα 5.3.1), επιλέξαμε την αξιολόγηση των μεθόδων και των κριτηρίων διαφοροποίησης, χρησιμοποιώντας πλήθος μετρικών που χρησιμοποιούνται στα TREC Diversity Tasks[16]. Συγκεκριμένα, χρησιμοποιούμε:

- a-nDCG: Η μετρική a-normalized discounted cumulative gain [30] ποσοτικοποιεί το ποσοστό μοναδικών πτυχών του ερωτήματος q που καλύπτονται από τα $top - k$ κατεταγμένα έγγραφα. Χρησιμοποιούμε $a = 0.5$, όπως συνήθως χρησιμοποιείται στην αξιολόγηση TREC.

- Precision-IA: Η μετρική Precision-Intent Aware [3] εκφράζει την αναλογία σχετικών εγγράφων για διάφορα επιμέρους θέματα στα $top - k$ κατεταγμένα έγγραφα.

- S-recall: Η μετρική Subtopic-recal [155] υπολογίζεται ως ο λόγος των μοναδικών πτυχών (aspects) που καλύπτονται από τα $top - k$ αποτελέσματα με τον συνολικό αριθμό των πτυχών. Μετρά σε συγκεκριμένη λίστα αποτελεσμάτων την κάλυψη μιας πτυχής σε βάθος k.

5.4.3 Κρίσεις Συνάφειας

Για την πειραματική αξιολόγηση των τεχνικών και των κριτηρίων διαφοροποίησης ακολουθήσαμε την μεθοδολογία που περιγράψαμε στην Ενότητα 5.3.3, αρχικά για τον καθορισμό των ερωτήσεων και εν συνεχεία για την υποσημείωση της συλλογής νομικών εγγράφων με κρίσεις συνάφειας για κάθε ερώτημα. Συνολικά επιλέξαμε 330 ερωτήματα που καλύπτουν 1,650 πτυχές/ θέματα. Διαθέσαμε ελεύθερα στο διαδίκτυο [17] το πλήρες σύνολο δεδομένων, τα δεδομένα αντικειμενικής αλήθειας, τα ερωτήματα και τις επισημειώσεις των εγγράφων για κάθε ερώτημα, με στόχο την ενίσχυση της συνεργασίας σε θέματα διαφοροποιημένης ανάκτησης νομικών κειμένων.

[16] `http://trec.nist.gov/data/web10.html`
[17] https://github.com/mkoniari/MultiLegalDiv

5.4.4 Αποτελέσματα

Ως βάση για την σύγκριση των μεθόδων διαφοροποίησης, θεωρούμε την αρχική κατάταξη των εγγράφων, που παράγεται από την συνάρτηση ομοιότητας εγγράφου με το ερώτημα του χρήστη (Εξίσωση 5.8) και το log-based $tf - idf$ σχήμα ευρετηρίασης. Για κάθε ερώτημα, το αρχικό σύνολο N περιέχει τα $top-n$ αποτελέσματα. Για όλες τις μεθόδους διαφοροποίησης η παράμετρος trade-off λ ρυθμίζεται σε σταθερή τιμή 0.5 και επομένως $\lambda = 0.5 = 1-\lambda$, δηλαδή επιδιώκουμε ισοκατανομή των βαρών της συνάφειας και της διαφορετικότητας στο τελικό αποτέλεσμα. Παρουσιάζουμε τα αποτελέσματα της αξιολόγησης για τα σενάρια - μεθόδους που εξετάζουμε, χρησιμοποιώντας τις προαναφερθείσες μετρικές αξιολόγησης, με τιμές cut-off @5, @10, @20, όπως είθισται στις αξιολογήσεις TREC. Τα αποτελέσματα παρουσιάζονται με σταθερή παράμετρο $n = |N|$. Σημειώνουμε ότι κάθε μία από τις δοκιμές διαφοροποίησης εφαρμόζεται σε συνδυασμό με κάθε ένα από τους αλγορίθμους διαφοροποίησης και για κάθε ερώτημα του χρήστη.

Ο Πίνακας 5.6 συνοψίζει τις παραμέτρους δοκιμών και τις αντίστοιχες τιμές τους. Κάθε ερώτημα χρήστη υποβάλλεται στο σύστημα και ανακτώνται τα $top-n$ έγγραφα, τα οποία σχηματίζουν το υποψήφιο σύνολο N, με χρήση της Εξίσωσης 5.5 και το log-based $tf - idf$ σχήμα ευρετηρίασης. Η επιλεγμένη τιμή, 100, για το πλήθος του υποψηφίου συνόλου N είναι μια τυπική τιμή που χρησιμοποιείται στη βιβλιογραφία [126]. Οι πειραματικές μελέτες μας πραγματοποιούνται σε μια διπλή στρατηγική: (α) ποιοτική ανάλυση από την άποψη της διαφοροποίησης και της ακρίβειας της κάθε μέθοδος που χρησιμοποιείται σε σχέση με το βέλτιστο σύνολο των αποτελεσμάτων και (β) ανάλυση της επεκτασιμότητας (scalability analysis) των μεθόδων διαφοροποίησης, όταν αυξάνουν οι τιμές των παραμέτρων.

Πίνακας 5.6: Παράμετροι Πειραματικής Μελέτης

Παράμετρος	**Εύρος Τιμών**
Αλγόριθμοι Αξιολόγησης	MMR, Max-min, Max-sum, Mono-objective, LexRank, Biased LexRank, DivRank, Grasshopper
tradeoff λ	0.5
Αριθμός Εγγράφων $n = \|N\|$	100
Αριθμός Αποτελεσμάτων $k = \|S\|$	5, 10, 20
Υποβληθέντα Ερωτήματα	330
Περίπτωση 1 Βάρη Κριτηρίων	Κειμενικό περιεχόμενο 1.0, Χρόνος, 0 Ποιότητα γραφής 0, Θεματικές Κατηγορίες, 0
Περίπτωση 2 Βάρη Κριτηρίων	Κειμενικό περιεχόμενο 0.6, Χρόνος, 0.13, Ποιότητα γραφής 0.13, Θεματικές Κατηγορίες, 0.14

Αρχικά χρησιμοποιήσαμε τις μεθόδους διαφοροποίησης κάνοντας χρήση μόνο του Κειμενικού περιεχόμενου, όπως είθισται στις περιπτώσεις διαφοροποίησης αποτελεσμάτων αναζήτησης. Στην περίπτωση αυτή τα βάρη των υπόλοιπων κριτηρίων (Χρόνος, Ποιότητα γραφής και Θεματικές Κατηγορίες), στην Εξίσωση Ομοιότητας Εγγράφων (Εξίσωση 5.6) τέθηκαν στο μηδέν. Ο Πίνακας 5.7 συνοψίζει το μέσο όρο (average) αποτελεσμάτων των υπό αξιολόγηση μεθόδων διαφοροποίησης. Στατιστικώς σημαντικές τιμές, χρησιμοποιώντας το paired

two-sided t-test, με τιμές $p_{value} < 0.05$ σημειώνονται με $^\circ$ και για $p_{value} < 0.01$ με *.

Οι μέθοδοι MMR και DivRank είναι οι καλύτερες στρατηγικές διαφοροποίησης για διάφορες μετρήσεις αξιολόγησης με $N = 100$ και $k = 30$. Ειδικότερα, η μέθοδος MMR ξεπερνά όλες τις άλλες μεθόδους όσον αφορά τις μετρικές a-nDCG και ST Recall σε όλα τα επίπεδα, ενώ η μέθοδος DivRank επιτυγχάνει την υψηλότερη βαθμολογία για την μετρική Precision IA σε όλα τα επίπεδα. Οι μέθοδοι περιλήψεων (LexRank, Biased LexRank) και η μέθοδος Grasshopper, όπως χρησιμοποιήθηκε χωρίς την χρήση του δικτύου νομικών παραπομπών, αποτυγχάνουν να βελτιώσουν την μέτρηση βάσης, επιφέροντας χειρότερα αποτελέσματα από την αρχική τιμή κατάταξης σε όλα τα επίπεδα για όλες τις μετρήσεις. Η μέθοδος MMR επιτυγχάνει τα καλύτερα αποτελέσματα σε σχέση με τις υπόλοιπες μεθόδους διαφοροποίησης αποτελεσμάτων αναζήτησης, με εξαίρεση το επίπεδο nDCG@5, όπου η μέθοδος Max-min έχει καλύτερα αποτελέσματα. Η μέθοδος διαφοροποίησης γράφων DivRank, ενώ γενικά δεν βελτιώνει την μέτρηση βάσης στις μετρικές a-nDCG και ST Recall, επιτυγχάνει την υψηλότερη βαθμολογία για την μετρική Precision IA σε όλα τα επίπεδα. Σημειώνουμε ότι η απόδοση της μεθόδου DivRank, εμφανίζεται αισθητά βελτιωμένη σε αυτήν την πειραματική μελέτη, σε σύγκριση με την μελέτη της Ενότητας 5.3. Η βελτίωση αυτή αποδίδεται στην συνεκτικότητα του γράφου νομικών παραπομπών που χαρακτηρίζει την τρέχουσα συλλογή νομικών εγγράφων, καθώς αυτή εκτείνεται σε χρονικό διάστημα πλέον των 2 αιώνων. Αντιθέτως στην περίπτωση της Ενότητας 5.3, η συλλογή νομικών εγγράφων εκτείνεται σε χρονικό διάστημα 3 ετών, με αποτέλεσμα ο γράφος νομικών παραπομπών να είναι ιδιαίτερα αραιός και η μέθοδος DivRank να επιφέρει χειρότερα αποτελέσματα από την αρχική τιμή κατάταξης.

Πίνακας 5.7: Επίδοση των αξιολογούμενων μεθόδων, χρησιμοποιώντας μόνο Κειμενικό περιεχόμενο για τιμές παραμέτρων $|N| = 100$, $k = 30$. Οι υψηλότερες βαθμολογίες εμφανίζονται με έντονους χαρακτήρες. Στατιστικά σημαντικές τιμές, χρησιμοποιώντας το paired two-sided t-test, με τιμές $p_{value} < 0.05$ σημειώνονται με $^\circ$ και για $p_{value} < 0.01$ με *.

	a-nDCG			Precision IA			ST Recall		
Αλγόριθμος	@5	@10	@20	@5	@10	@20	@5	@10	@20
baseline	0,532	0,595	0,656	0,314	0,313	0,314	0,688	0,833	0,948
MMR	**0,571***	**0,643***	**0,695***	0,315°	0,321	0,322*	**0,783***	**0,923***	**0,977***
Max-sum	0,549	0,620*	0,675*	0,300*	0,305°	0,303*	0,744*	0,880*	0,969°
Max-min	0,568*	0,633*	0,686*	0,319	0,319*	0,319*	0,777*	0,907*	0,976*
Mono-objective	0,541°	0,602°	0,664*	0,313	0,310	0,312	0,713*	0,844	0,960°
LexRank	0,487	0,532*	0,586*	0,308*	0,313	0,320	0,584*	0,705*	0,820*
Biased LexRank	0,488*	0,533*	0,587*	0,309	0,314	0,320	0,585*	0,708*	0,821*
DivRank	0,533	0,589	0,635	**0,320**	**0,326**$^\circ$	**0,326**$^\circ$	0,667	0,803	0,888*
Grasshopper	0,492*	0,542*	0,598*	0,310	0,316	0,322°	0,592*	0,725*	0,846*

Στην συνέχεια αξιολογούμε τις μεθόδους διαφοροποίησης κάνοντας χρήση όλων των κριτηρίων διαφοροποίησης νομικών εγγράφων που προτείνουμε. Ειδικότερα, στην Εξίσωση Ομοιότητας Εγγράφων (5.5) οι τιμές των κριτηρίων διαφοροποίησης τίθενται σε: Κειμενικό περιεχόμενο 0.6, Χρόνος 0.13, Ποιότητα γραφής 0.13 ανδ Θεματικές Κατηγορίες 0.14. Σημειώνουμε ότι τα κριτήρια αυτά δεν μπορούν να εφαρμοστούν στην περίπτωση της μεθόδου DivRank, όπου χρησιμοποιείται ο γράφος νομικών παραπομπών των αποτελεσμάτων αναζήτησης για

κάθε ερώτημα χρήστη. Ο Πίνακας 5.8 συνοψίζει το μέσο όρο (average) αποτελεσμάτων των υπό αξιολόγηση μεθόδων διαφοροποίησης. Στατιστικά σημαντικές τιμές, χρησιμοποιώντας το paired two-sided t-test, με τιμές $p_{value} < 0.05$ σημειώνονται με $^\circ$ και για $p_{value} < 0.01$ με *.

Πίνακας 5.8: Επίδοση των αξιολογούμενων μεθόδων, χρησιμοποιώντας όλα τα κριτήρια διαφοροποίησης για τιμές παραμέτρων $|N| = 100$, $k = 30$. Οι υψηλότερες βαθμολογίες εμφανίζονται με έντονους χαρακτήρες. Στατιστικά σημαντικές τιμές, χρησιμοποιώντας το paired two-sided t-test, με τιμές $p_{value} < 0.05$ σημειώνονται με $^\circ$ και για $p_{value} < 0.01$ με *.

	a-nDCG			Precision IA			ST Recall		
Αλγόριθμος	@5	@10	@20	@5	@10	@20	@5	@10	@20
baseline	0,532	0,595	0,656	0,314	0,313	0,314	0,688	0,833	0,948
MMR	**0,586***	**0,657***	**0,709***	0,321	0,321°	0,325*	**0,815***	**0,939***	**0,989***
Max-sum	0,564*	0,636*	0,689*	0,306	0,306°	0,308°	0,779*	0,913*	0,977*
Max-min	0,581*	0,650*	0,702*	**0,322**	0,322	0,321*	0,793*	0,931*	0,983*
Mono-objective	0,550*	0,612*	0,673*	0,321°	0,313	0,314	0,716*	0,857°	0,968*
LexRank	0,484*	0,532	0,587*	0,304*	0,306	0,316	0,604*	0,724*	0,839*
Biased LexRank	0,488*	0,537*	0,592*	0,304	0,308	0,316	0,607*	0,731*	0,845*
DivRank	0,533	0,589	0,635	0,320	**0,326***	**0,326**$^\circ$	0,667	0,803	0,888*
Grasshopper	0,504°	0,555*	0,612*	0,306	0,308	0,317	0,649	0,760*	0,880*

Συνολικά αποδεικνύεται ότι η υιοθέτηση περισσότερο εξειδικευμένων κριτηρίων από την απλή κειμενική ομοιότητα μπορεί να βελτιώσει την αποτελεσματικότητα της διαδικασίας διαφοροποίησης. Επιπλέον οι τεχνικές διαφοροποίησης αποτελεσμάτων αναζήτησης στο διαδίκτυο ξεπερνούν τις άλλες προσεγγίσεις (π.χ., μέθοδοι περιληπτική παρουσίαση που βασίζεται σε γράφημα), στο πλαίσιο των νομικών διαφοροποίησης αναζήτησης. Η διαφοροποίηση με βάση το δίκτυο νομικών αναφορών, DivRank, αποτυγχάνει να βελτιώσει την μέτρηση βάσης στις μετρικές a-nDCG και S-recall, αλλά ξεπερνά τις υπόλοιπες μεθόδους στην μετρική Precision IA.

5.5 Συμπεράσματα

Στο κεφάλαιο αυτό μελετήθηκε το πρόβλημα της διαφοροποιημένης ανάκτησης νομικών πηγών. Εισάγαμε εξειδικευμένα κριτήρια διαφοροποίησης αναλύοντας την επίδραση διαφόρων χαρακτηριστικών των νομικών πηγών στον υπολογισμό της ομοιότητας των εγγράφων με το ερώτημα του χρήστη και των εγγράφων μεταξύ τους. Προσαρμόσαμε στο συγκεκριμένο πρόβλημα και στα κριτήρια που εισάγαμε διαδεδομένους αλγόριθμους που έχουν προταθεί για την κάλυψη διαφορετικών αναγκών π.χ. την δημιουργία περιλήψεων, την κατάταξη σε γράφους, παράλληλα με αλγόριθμους διαφοροποίησης αποτελεσμάτων αναζήτησης. Οι προαναφερθείσες μέθοδοι και κριτήρια, αξιολογήθηκαν, με βάση διεθνώς αποδεκτές μετρικές, σε πραγματικές συλλογές νομικών εγγράφων προερχόμενες από διαφορετικά νομικά συστήματα.

Η πειραματική αξιολόγηση έδειξε ότι η υιοθέτηση μεθόδων διαφοροποίησης, στο πλαίσιο ανάκτησης νομικής πληροφορίας, επιφέρει στατιστικά αξιοσημείωτες βελτιώσεις όσον αφορά τον εμπλουτισμό των αποτελεσμάτων αναζήτησης με κατά τα άλλα κρυφές πτυχές του νομικού χώρου γύρω από το ερώτημα του χρήστη καθώς και ότι τα κριτήρια διαφοροποίησης που προ-

τείνουμε επίσης παρέχουν στατιστικά αξιοσημείωτα διαφοροποιημένα υποσύνολα ανακτημένων νομικών εγγράφων. Επιπρόσθετα οι μέθοδοι κατάταξης σε γράφους επιφέρουν αξιοσημείωτα αποτελέσματα μόνο σε περιπτώσεις πληρότητας του σώματος νομικών πηγών, ενώ οι μέθοδοι διαφοροποίησης αποτελεσμάτων αναζήτησης ξεπερνούν, στο πλαίσιο της νομικής διαφοροποίησης, τις μεθόδους δημιουργίας περιλήψεων κειμένων. Επιπρόσθετα η αξιολόγηση που πραγματοποιήθηκε, για τις ανάγκες της οποίας αναπτύξαμε ειδική μεθοδολογία αντικειμενικής επισημείωσης του συνόλου δεδομένων, προσφέρει όρια εξισορρόπησης για τα συστήματα ανάκτησης νομικής πληροφορίας, που επιθυμούν να ισορροπήσουν μεταξύ της ενίσχυσης των σχετικών εγγράφων ή να δειγματοληπτήσουν τον χώρο νομικής πληροφορίας γύρω από το ερώτημα.

Σε μελλοντικές εργασίες, σχεδιάζουμε να μελετήσουμε περαιτέρω την αλληλεπίδραση της σχετικότητας και της ποικιλομορφίας σε ιστορικά νομικά ερωτήματα. Ενώ η πρόσβαση στις νομικές πηγές γενικά ανακτά την ισχύουσα νομοθεσία σε ένα θέμα, τα νομικά συστήματα πληροφοριών τύπου point-in-time λειτουργούν με διαφορετική οπτική: οι δικηγόροι, οι δικαστές, αλλά και απλοί πολίτες, πολλές φορές αξιολογώντας τις νομικές συνέπειες των γεγονότων του παρελθόντος, θα πρέπει να γνωρίζουν την ισχύουσα νομοθεσία σε κάποιο σημείο του παρελθόντος, όταν συνέβησαν γεγονότα που οδήγησαν σε διένεξη ή σε δικαστικές διαμάχες [153].

Κεφάλαιο 6

Διαφοροποιημένη ανάκτηση καταχωρήσεων σε κείμενα διαβουλεύσεων και σε κοινωνικά δίκτυα

Σε αυτό το κεφάλαιο παρουσιάζουμε τις μεθόδους που προτείνουμε για διαφοροποιημένη ανάκτηση καταχωρήσεων σε κείμενα διαβουλεύσεων και κοινωνικά δίκτυα. Η προτεινόμενη μέθοδος αξιολογήθηκε για την διαφοροποίηση καταχωρήσεων χρηστών σε ειδησεογραφικά άρθρα [54] και για την διαφοροποίηση καταχωρήσεων χρηστών σε κοινωνικά δίκτυα [76].

6.1 Διαφοροποιημένη Ανάκτηση σε κείμενα διαβουλεύσεων

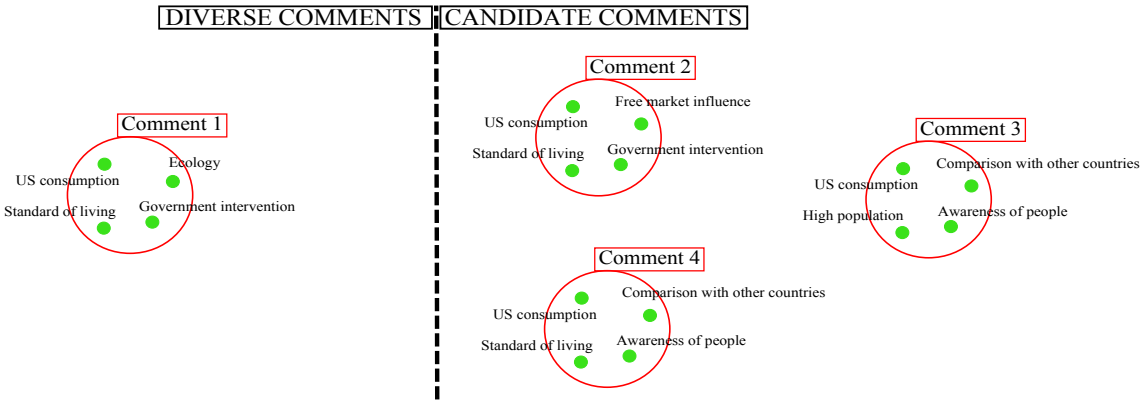
Σε αυτήν την ενότητα αρχικά παρουσιάζουμε, μέσω παραδείγματος, την χρησιμότητα της προτεινόμενης μεθόδου, καταδεικνύοντας ταυτόχρονα την ανάγκη επέκτασης των τεχνικών διαφοροποίησης για την αντιμετώπιση των ιδιαίτερων απαιτήσεων του προβλήματος. Στην συνεχεία ορίζουμε το πρόβλημα και παρουσιάζουμε αναλυτικά τα προτεινόμενα κριτήρια διαφοροποίησης. Προσαρμόζουμε στα κριτήρια που εισάγουμε διαδεδομένους ευρετικούς αλγορίθμους και προτείνουμε επίσης μια παραλλαγή ενός αλγορίθμου. Τέλος, πραγματοποιούμε εκτενή πειραματική αξιολόγηση των μεθόδων και κριτηρίων διαφοροποίησης που εισάγουμε.

6.1.1 Κίνητρο και Συνεισφορά

Θεωρούμε ως μονάδα πληροφορίας για σχόλια χρηστών κάθε έννοια ή θεματική κατηγορία που σχετίζεται με το κυρίως θέμα/άρθρο ή κάθε σχετική άποψη/συναίσθημα ή επέκταση των παραπάνω εννοιών και κατηγοριών, τα οποία μπορεί να εντοπιστούν, είτε στο κυρίως άρθρο, είτε στα σχόλια του. Στόχος της διαδικασίας της διαφοροποίησης είναι να συγκεντρώσει ένα

υποσύνολο σχολίων το οποίο θα περιέχει όσο το δυνατόν περισσότερες και πιο ετερογενείς μονάδες πληροφορίας που να αφορούν το κυρίως άρθρο.

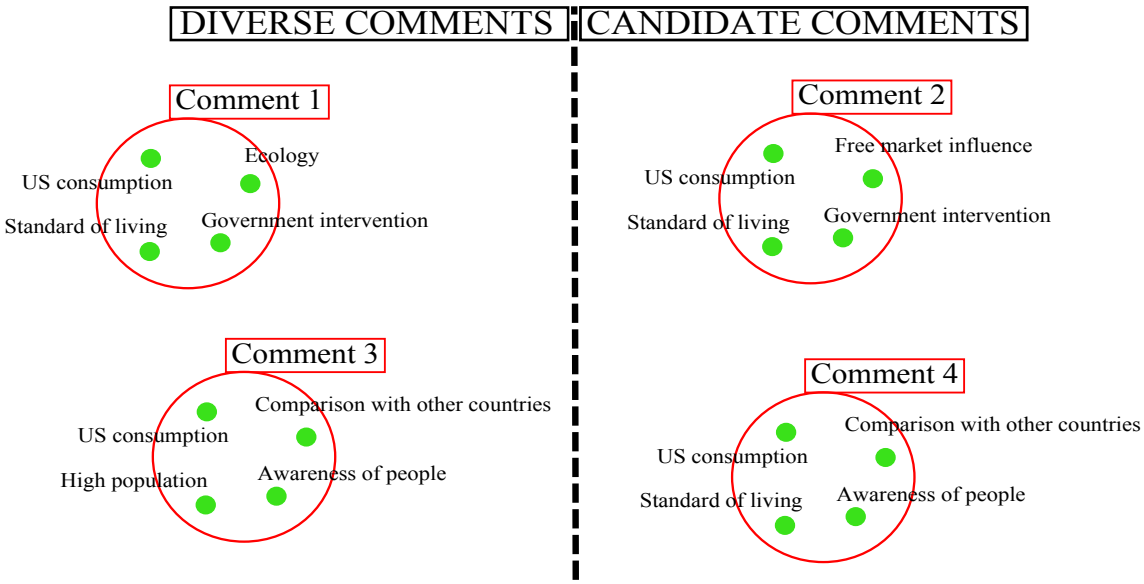
Στα σχήματα που ακολουθούν (Σχήμα 6.1, Σχήμα 6.2) παρουσιάζουμε ένα παράδειγμα διαφοροποίησης αναρτήσεων χρηστών με βάση τις μονάδες πληροφορίας (information nuggets). Το Σχήμα 6.1 παρουσιάζει τέσσερα σχόλια για ένα άρθρο με θέμα "US consumption". Υποθέτουμε επίσης, για απλούστευση, χωρίς βλάβη της γενικότητας, ότι κάθε σχόλιο περιέχει τέσσερις διαφορετικές μονάδες πληροφορίας σχετιζόμενες με το άρθρο. Υποθέτουμε ότι έχει ήδη επιλεγεί το σχόλιο 1, ως το πρώτο αποτέλεσμα του διαφοροποιημένου συνόλου αποτελεσμάτων. Σε αυτήν την περίπτωση, στόχος της διαδικασίας διαφοροποίησης είναι να επιλέξει ένα από τα εναπομείναντα υποψήφια σχόλια, μεγιστοποιώντας την ετερογένεια των μονάδων πληροφορίας στο σύνολο αποτελεσμάτων που θα προκύψει. Από τα υποψήφια σχόλια, το σχόλιο 2 έχει 1/4 μονάδες πληροφορίας διαφορετικές από το ήδη επιλεγμένο σχόλιο 1, το 3 έχει 3/4 μονάδες διαφορετικές από το 1 και το σχόλιο 4 έχει 2/4 διαφορετικές μονάδες από το 1. Επομένως, το σχόλιο 3 είναι αυτό που επιλέγεται ως το επόμενο αποτέλεσμα για τον σχηματισμό ενός διαφοροποιημένου συνόλου αποτελεσμάτων.



Σχήμα 6.1: Διαδικασία Διαφοροποιημένης Ανάκτησης Σχολίων Χρηστών - (α)

Στην συνέχεια, Σχήμα 6.2, δεδομένου ότι το σύνολο αποτελεσμάτων περιέχει τα σχόλια 1 και 3, το πιο διαφοροποιημένο υποψήφιο σχόλιο είναι το 2, αφού περιέχει τη μονάδα πληροφορίας Freemarket influence η οποία δεν περιέχεται ούτε στο σχόλιο 1, ούτε στο 3. Ταυτόχρονα, όλες οι μονάδες του σχολίου 4 περιέχονται στα σχόλια του συνόλου αποτελεσμάτων, οπότε το σχόλιο 2 είναι αυτό που επιλέγεται ως το επόμενο αποτέλεσμα.

Το παράδειγμα αυτό καταδεικνύει πώς η διαφοροποίηση/ετερογένεια ενός συνόλου σχολίων όσον αφορά σε ένα άρθρο, μπορεί να ποσοτικοποιηθεί και να αξιολογηθεί μέσω των μονάδων πληροφορίας που εξάγονται από το άρθρο και τα σχόλιά του. Ταυτόχρονα, η αναγνώριση εννοιών, θεμάτων και γνωμών, είναι πολύ δύσκολη εργασία, ακόμα και όταν γίνεται από κάποιον κριτή. Ακόμα και δημοφιλή εργαλεία εξαγωγής οντοτήτων, θεμάτων ή συναισθημάτων δεν μπορούν να αναγνωρίσουν όλες τις έννοιες που περιέχονται σε ένα κομμάτι κειμένου ή να αναγνωρίσουν με συνέπεια την ίδια οντότητα, η οποία περιέχεται αυτούσια σε διαφορετικά κομμάτια κειμένου ή με ελαφρά διαφορετικές μορφές. Για παράδειγμα, όταν οι χρήστες χρησιμοποιούν εκφράσεις όπως "Republicans", "conservatives" ή "Bushs gover-



Σχήμα 6.2: Διαδικασία Διαφοροποιημένης Ανάκτησης Σχολίων Χρηστών - (β)

nance", ανάλογα με τα συμφραζόμενα του άρθρου, υπάρχει περίπτωση να αναφέρονται στην ίδια ή σε διαφορετικές έννοιες. Καθώς η ακριβής διάκριση των εννοιών είναι αδύνατη, τα εξειδικευμένα σε σχόλια χρηστών κριτήρια διαφοροποίησης που προτείνουμε, προσπαθούν, με έμμεσο αλλά αυτοματοποιημένο τρόπο, να αιχμαλωτίσουν τις πτυχές των εννοιών στα σχόλια των χρηστών.

Τα κριτήρια που προτείνουμε μοντελοποιούν τα κάτωθι χαρακτηριστικά των σχολίων:

- Κειμενικό περιεχόμενο. Στόχος είναι η ανάκτηση σχολίων με όσο το δυνατόν πιο ετερογενές κειμενικό περιεχόμενο.
- Συναίσθημα. Στόχος είναι η ανάκτηση σχολίων που καλύπτουν ένα ανομοιογενές φάσμα συναισθημάτων που εκφράζουν οι χρήστες στα σχόλιά τους.
- Ονοματικές οντότητες. Στόχος είναι το διαφοροποιημένο σύνολο σχολίων να περιέχει όσο το δυνατόν περισσότερες/διαφορετικές ονοματικές οντότητες (Πρόσωπα, Οργανισμοί, Τοποθεσίες) που εντοπίζονται στο άρθρο και στα σχόλια
- Συναίσθημα γύρω από ονοματικές οντότητες. Στόχος είναι η ανάκτηση σχολίων που καλύπτουν ένα ανομοιογενές φάσμα συναισθημάτων για τις ονοματικές οντότητες που εκφράζουν οι χρήστες στα σχόλιά τους.
- Ποιότητα γραφής σχολίου. Στόχος είναι η συλλογή σχολίων διαφορετικής ποιότητας γραφής.
- Αθροιστική ποιότητα γραφής σχολίου. Παράγοντας για κάθε χρήστη, μία μέση ποιότητα γραφής, με βάση όλα τα σχόλια που έχει κάνει συνολικά για όλα τα άρθρα, στόχος είναι η συλλογή σχολίων διαφορετικής αθροιστικής ποιότητας γραφής.

Στην συνέχεια, ενσωματώνουμε τα κριτήρια διαφοροποίησης που εισάγουμε, σε διαδεδομένους ευρετικούς αλγόριθμους για την διαφοροποίηση αποτελεσμάτων αναζήτησης. Παράλληλα, προτείνουμε μια παραλλαγή ενός αλγορίθμου και πραγματοποιούμε εκτενή πειραματική αξιολόγηση των μεθόδων και κριτηρίων διαφοροποίησης που εισάγουμε.

Η αξιολόγηση γίνεται με τη βοήθεια τριών μετρικών αξιολόγησης που ορίζουμε, με σκοπό να ποσοτικοποιήσουμε τον αριθμό των μονάδων πληροφορίας που απαντώνται σε κάθε διαφοροποιημένο σύνολο αποτελεσμάτων, καθώς επίσης και την ομοιογένεια των μονάδων πληροφορίας στα διαφορετικά σύνολα σχολίων. Τα αποτελέσματα της αξιολόγησης φανερώνουν την αποτελεσματικότητα των μεθόδων μας, ενάντια στη βασική μέθοδο της διαφοροποίησης σχολίων μόνο με βάση το κειμενικό περιεχόμενο. Σχεδόν όλες οι παραλλαγές της μεθόδου μας αποδίδουν καλύτερα από τη βασική μέθοδο όσον αφορά στην κάλυψη διαφορετικών μονάδων πληροφορίας, με την καλύτερη παραλλαγή να έχει στατιστικά σημαντικά καλύτερη αποτελεσματικότητα.

Συνολικά, η συνεισφορά της εργασίας μας είναι η ακόλουθη:

1. Ορίζουμε εξειδικευμένα σε σχόλια χρηστών κριτήρια διαφοροποίησης
2. Προτείνουμε μία παραλλαγή ευριστικού αλγόριθμου διαφοροποίησης που αποδίδει πολύ κοντά στον βέλτιστο δοκιμαζόμενο αλγόριθμο
3. Επεκτείνουμε την έννοια των μονάδων πληροφορίας στο σενάριο των άρθρων/σχολίων, ορίζοντας μετρικές για να αξιολογήσουμε την αποτελεσματικότητα των μεθόδων,
4. Πραγματοποιούμε μία αναλυτική αξιολόγηση, χρησιμοποιώντας δημοσίως διαθέσιμα σύνολα δεδομένων από άρθρα και σχόλια, αποδεικνύοντας την αποτελεσματικότητα των μεθόδων μας.

6.1.2 Ορισμός Προβλήματος

Στην ενότητα αυτή μελετάμε το πρόβλημα της διαφοροποίησης σχολίων ειδησεογραφικών άρθρων. Έχοντας ένα ειδησεογραφικό άρθρο και μια συλλογή από σχόλια χρηστών για το άρθρο αυτό, στόχος της διαδικασίας διαφοροποίησης είναι να συγκεντρώσει ένα υποσύνολο σχετικών και αντιπροσωπευτικών με το άρθρο σχολίων και να επιλέξει τα σχόλια αυτά με τέτοιο τρόπο ώστε η ποικιλομορφία του συνόλου να μεγιστοποιείται. Πιο συγκεκριμένα, το πρόβλημα ορίζεται ως εξής:

Ορισμός 6.1 (Διαφοροποίηση Σχολίων Χρηστών)**.** *Έστω A ένα άρθρο και N ένα σύνολο από σχόλια χρηστών στο άρθρο. Βρείτε ένα υποσύνολο $S \subset N$ σχολίων που μεγιστοποιεί μία συνάρτηση στόχο f η οποία ποσοτικοποιεί την ετερογένεια των σχολίων στο S.*

Οι πιο πρόσφατες εργασίες στη διαφοροποίηση καταπιάνονται με τη διαφοροποίηση αποτελεσμάτων αναζήτησης, όσον αφορά το αντίστοιχο ερώτημα. Το δικό μας σενάριο διαφοροποίησης έχει αρκετές ομοιότητες με τη διαφοροποίηση αποτελεσμάτων αναζήτησης, για παράδειγμα, το γεγονός ότι, και στις δύο περιπτώσεις, τα στοιχεία προς διαφοροποίηση έχουν

κειμενικές περιγραφές. Επίσης, και στις δύο περιπτώσεις, πέρα από την ετερογένεια, πρέπει να ληφθεί υπόψη και η ομοιότητα των προς διαφοροποίηση στοιχείων (αποτελέσματα/σχόλια) με τον αντίστοιχο βασικό πόρο. Παρόλα αυτά, υπάρχουν και ουσιώδεις διαφορές που επιβάλλουν την ανάγκη ανάλυσης και επέκτασης/προσαρμογής των αλγορίθμων/κριτηρίων διαφοροποίησης, ειδικά για το σενάριο της διαφοροποίησης σχολίων. Στη συνέχεια αναλύουμε εν συντομία αυτές τις διαφορές που θεωρούμε πιο κρίσιμες:

- Βασικός πόρος. Τα ερωτήματα των χρηστών σε ένα σύστημα αναζήτησης είναι σύντομα σε έκταση και τις περισσότερες φορές αντιπροσωπεύουν μία ή λίγες ανάγκες αναζήτησης, οι οποίες είναι στενά συνδεδεμένες με τις λέξεις του ερωτήματος. Έτσι, στο σενάριο της αναζήτησης η διαφοροποίηση στοχεύει κυρίως στο διαχωρισμό διαφορετικών εκφάνσεων των όρων του ερωτήματος και στην παρουσίαση αποτελεσμάτων που καλύπτουν καλύτερα αυτές τις εκφάνσεις. Από την άλλη πλευρά, ένα άρθρο περιέχει πολύ περισσότερο κείμενο. Ειδικότερα, αποτελείται από μία ολοκληρωμένη περιγραφή ενός ή περισσοτέρων θεμάτων, τα οποία μπορεί να αναφέρονται σε επιμέρους θέματα. Επομένως, ο αριθμός των εννοιών προς διαφοροποίηση διαφέρει σε σχέση με την αναζήτηση αποτελεσμάτων. Ταυτόχρονα, οι οντότητες προς διαφοροποίηση μπορεί να μην είναι συμπαγείς έννοιες, αλλά να περιέχουν υπό-έννοιες, όπως στα Σχήματα 6.1 και 6.2

- Στοιχεία προς διαφοροποίηση. Τα αποτελέσματα αναζήτησης που επιστρέφονται από τις μηχανές αναζήτησης είναι σε κάποιο βαθμό σχετικά με το αντίστοιχο ερώτημα. Τα περισσότερα από αυτά αναμένεται να περιέχουν καλά δομημένο κείμενο που περιγράφει με σαφήνεια ένα ή περισσότερα θέματα σχετικά με το ερώτημα, καθώς η ποιότητα αυτών των αποτελεσμάτων έχει εξακριβωθεί από τους μηχανισμούς ταξινόμησης των μηχανών αναζήτησης. Από την άλλη πλευρά, τα σχόλια χρηστών συνήθως περιέχουν πολύ λιγότερο κείμενο και ετερογενή ποιότητα π.χ., έλλειψη σημείων στίξης, συντομογραφίες κ.λπ. Επίσης, μερικά σχόλια μπορεί να είναι απαντήσεις ή συνέχειες προηγούμενων σχολίων.

- Απόψεις χρηστών. Τις περισσότερες φορές τα αποτελέσματα αναζήτησης περιέχουν περιγραφές εννοιών που σχετίζονται με τους όρους του ερωτήματος. Σε αντίθεση, τις περισσότερες φορές τα σχόλια εκφράζουν απόψεις και συναισθήματα σχετικά με τις προς συζήτηση έννοιες.

- Ονοματικές οντότητες. Στο σενάριο των άρθρων/σχολίων, οι ονοματικές οντότητες διαδραματίζουν σημαντικό ρόλο καθώς αρκετές από τις έννοιες που περιγράφονται σε ένα άρθρο αναμένεται να σχετίζονται με ονοματικές οντότητες, ενώ και οι χρήστες σχολιάζουν ένα θέμα αναφερόμενοι σε ονοματικές οντότητες.

6.1.3 Κριτήρια Διαφοροποίησης

Στην ενότητα αυτή παρουσιάζουμε τα κριτήρια που χρησιμοποιούμε και το τρόπο υπολογισμού τους.

– **Συναίσθημα**. Οι απόψεις των χρηστών πάνω στα θέματα του άρθρου εμπεριέχουν διαφορετικές εκφάνσεις συναισθημάτων (θετικό, αρνητικό ή ουδέτερο). Συνεπώς η ανάκτηση ενός συνόλου σχολίων που καλύπτουν, στο μέτρο του δυνατού με ομοιόμορφο τρόπο, διαφορετικές διαβαθμίσεις συναισθήματος ευνοεί την ετερογένεια. Ορίζουμε εννιά κλάσεις συναισθήματος μέσα στο διάστημα $[-4, 4]$, με την τιμή -4 να υποδηλώνει πολύ αρνητικό συναίσθημα, 0 ουδέτερο συναίσθημα και 4 πολύ θετικό. Η εξαγωγή του συναισθήματος γίνεται με χρήση του Sentistrength [139], το οποίο επιστρέφει τιμές συναισθήματος στο διάστημα $[-4, 4]$. Παρόλα αυτά μπορεί να εφαρμοστεί οποιοσδήποτε εργαλείο εξαγωγής συναισθήματος και αριθμός από διαβαθμίσεις συναισθήματος, χωρίς να αλλάξει η λογική του κριτηρίου. Σε κάθε σχόλιο αντιστοιχίζουμε δύο κατηγορίες τιμών συναισθήματος:

  (i) Μέγιστο/Ελάχιστο συναίσθημα σε όλο το κείμενο του σχολίου.

  (ii) Μέσος Όρος. Θεωρούμε το συναίσθημα κάθε πρότασης του σχολίου και εν συνεχεία σχηματίζουμε τον μέσο όρο.

  Η εξαγωγή συναισθήματος βασίζεται σε συγκεκριμένες λέξεις που εντοπίζονται στο κείμενο και εκφράζουν θετικό ή αρνητικό συναίσθημα. Με τους δύο διαφορετικούς τρόπους εξαγωγής συναισθήματος, μπορούμε να αναγνωρίσουμε διαφορετικές όψεις του εκφραζόμενου συναισθήματος. Για παράδειγμα, ένα σχόλιο μπορεί να περιέχει μόνο μία πρόταση με πολύ θετικό συναίσθημα σχετιζόμενο με ένα θέμα του άρθρου, ενώ το υπόλοιπο σχόλιο μπορεί να είναι συνολικά αρνητικό προς το σύνολο του άρθρου. Με τον τρόπο αυτό μπορούμε να εντοπίσουμε αυτή τη διαφορά, μέσω της εξαγωγής μέσου συναισθήματος. Με βάση τις τιμές συναισθήματος, για κάθε σχόλιο κατασκευάζουμε δύο διανύσματα μεγέθους εννιά χαρακτηριστικών, όπου το κάθε χαρακτηριστικό αντιστοιχεί σε μία βαθμίδα συναισθήματος και έχει τιμή 1 ή 0, ανάλογα με το αν το αντίστοιχο συναίσθημα έχει εντοπιστεί ή όχι.

– **Ονοματικές Οντότητες**. Τα ειδησεογραφικά άρθρα περιστρέφονται, τις περισσότερες φορές, γύρω από ονοματικές οντότητες. Τα πρόσωπα, οι οργανισμοί καθώς και οι τοποθεσίες που απαντώνται σε ένα άρθρο ή/και στα σχόλια αυτού αποτελούν συνήθως τους πρωταγωνιστές του άρθρου. Επομένως ένα διαφοροποιημένο σύνολο σχολίων οφείλει να περιλαμβάνει όσο το δυνατόν περισσότερες διαφορετικές ονοματικές οντότητες. Για κάθε μία από τις τρεις παραπάνω κατηγορίες ονοματικών οντοτήτων, ορίζουμε διανύσματα χαρακτηριστικών, με κάθε χαρακτηριστικό να αντιστοιχεί σε μία διακριτή ονοματική οντότητα που υπάρχει στο άρθρο ή στα σχόλιά του. Για κάθε σχόλιο, οι τιμές του κάθε χαρακτηριστικού του διανύσματος είναι οι συχνότητες των αντιστοίχων ονοματικών οντοτήτων μέσα σε αυτό. Επιπλέον, θεωρούμε ένα συγκεντρωτικό διάνυσμα που περιέχει όλες τις ονοματικές οντότητες, από όλες τις κατηγορίες, ως χαρακτηριστικά. Επομένως, συνολικά, ορίζουμε τέσσερα διανύσματα που αντιπροσωπεύουν τις ονοματικές οντότητες για κάθε σχόλιο.

– **Συναίσθημα στις Ονοματικές Οντότητες**. Οι απόψεις του συντάκτη του κειμένου και των χρηστών για τις ονοματικές οντότητες αποκαλύπτονται καλύτερα με την εξαγωγή του σχετικού συναισθήματος για κάθε ονοματική οντότητα. Το διαφοροποιημένο σύνολο σχολίων με παρόμοια λογική θα πρέπει να συγκεντρώνει ετερογενή συναισθήματα για κάθε ονοματική οντότητα. Για κάθε ονοματική οντότητα, που εντοπίστηκε στο άρθρο, ή/και στα σχόλια, θεωρούμε ένα παράθυρο ±5 λέξεων γύρω από αυτήν και εξάγουμε το συναίσθημα μόνο για τη συγκεκριμένη περιοχή. Στη συνέχεια, επεκτείνουμε το διάνυσμα ονοματικών οντοτήτων και σχηματίζουμε ένα νέο διάνυσμα όπου στην θέση κάθε χαρακτηριστικού ονοματικής οντότητας υπάρχουν εννιά χαρακτηριστικά, ένα για κάθε κλάση συναισθήματος.

– **Ποιότητα γραφής**. Η ποιότητα γραφής εκφράζει το βαθμό κατανόησης ενός σχολίου και συνεπώς ένα ετερογενές σύνολο σχολίων οφείλει να περιλαμβάνει και διαφορετικά επίπεδα καταληπτότητας - αναγνωσιμότητας. Η αναγνωσιμότητα ενός σχολίου υποδηλώνει το επίπεδο δυσκολίας του γραπτού κειμένου και προκύπτει από την εφαρμογή ενός τύπου αναγνωσιμότητας που λαμβάνει υπόψη ποιοτικά χαρακτηριστικά ενός κειμένου, όπως μήκος λέξεων και προτάσεων. Μια από τις πιο διαδεδομένες συναρτήσεις βαθμολόγησης της αναγνωσιμότητας δοθέντος κειμένου είναι η Flesch Reading Ease Score[1], η οποία συνδυάζει μέσο αριθμό συλλαβών ανά λέξη και μέσο μήκος πρότασης για να παράγει μία βαθμολογία καταληπτότητας. Συγκεκριμένα, παράγει μια τιμή στο διάστημα [0,100], με υψηλότερες τιμές να υποδεικνύουν ευκολότερα κείμενα. Σε αυτό το διάστημα ορίζουμε επτά κλάσεις αναγνωσιμότητας. Για κάθε σχόλιο εφαρμόζουμε τη συνάρτηση βαθμολόγησης και αναθέτουμε μία κλάση αναγνωσιμότητας σε αυτό. Στη συνέχεια, κατασκευάζουμε ένα διάνυσμα επτά χαρακτηριστικών και αναθέτουμε τιμή 1 στο χαρακτηριστικό που αντιπροσωπεύει την αναγνωσιμότητα του σχολίου και 0 στα υπόλοιπα χαρακτηριστικά.

– **Συναθροιστική Ποιότητα γραφής**. Αντίστοιχα με την ποιότητα γραφής, σχηματίζουμε τον μέσο όρο ποιότητας γραφής για όλα τα σχόλια του κάθε χρήστη. Με την τιμή αυτή χαρακτηρίζουμε πλέον κάθε σχόλιο του χρήστη.

– **Κειμενικό περιεχόμενο**. Τέλος, εξετάζουμε το περιεχόμενο των σχολίων, το οποίο αποτελεί το βασικό κριτήριο διαφοροποίησης, που χρησιμοποιείται στις περισσότερες εργασίες που σχετίζονται με την διαφοροποίηση. Η σημασία του περιεχομένου των σχολίων στη διαδικασία διαφοροποίησης είναι προφανής. Για κάθε σχόλιο, κατασκευάζουμε το διάνυσμα όρων του, με κάθε στοιχείο του διανύσματος να αντιστοιχεί σε κάθε διακριτό όρο που υπάρχει σε όλο το σώμα άρθρων/σχολίων. Η τιμή του κάθε στοιχείου του διανύσματος υπολογίζεται κανονικοποιώντας τη συχνότητα του αντίστοιχου όρου μέσα στο σχόλιο, με βάση το συνολικό αριθμό όρων του σχολίου.

[1] `http://en.wikipedia.org/wiki/Readability`

6.1.4 Αλγόριθμοι Διαφοροποίησης

Οι αλγόριθμοι διαφοροποίησης που υλοποιήσαμε, (Max-sum, Max-min, Mono-objective), έχουν παρουσιαστεί σε προηγούμενη εργασία [58] και περιγράφηκαν στην Ενότητα 5.2.4. Επιπρόσθετα στην παρούσα εργασία, προτείνουμε και υλοποιούμε τον κάτωθι αλγόριθμο, Max-sum2 Αλγόριθμος 6.1, για την συνάρτηση στόχο Max-sum. Σημειώνουμε ότι ενώ παραλλαγές της γενικής λογικής του αλγορίθμου μπορεί να έχουν προταθεί και σε άλλες δουλειές [3], υπάρχουν διαφορές τόσο στην αρχικοποίηση του αλγορίθμου όσο και στον ακριβή ορισμό της συνάρτησης απόστασης - ομοιότητας.

Στον αλγόριθμο Max-sum2 το διαφοροποιημένο σύνολο αποτελεσμάτων S αρχικοποιείται με το σχόλιο με την μεγαλύτερη κειμενική ομοιότητα με το άρθρο. Στην συνέχεια σε κάθε βήμα του, ο αλγόριθμος επιλέγει το υποψήφιο σχόλιο με την μεγαλύτερη απόσταση από το centroid του S. Το σχόλιο αυτό μεγιστοποιεί την

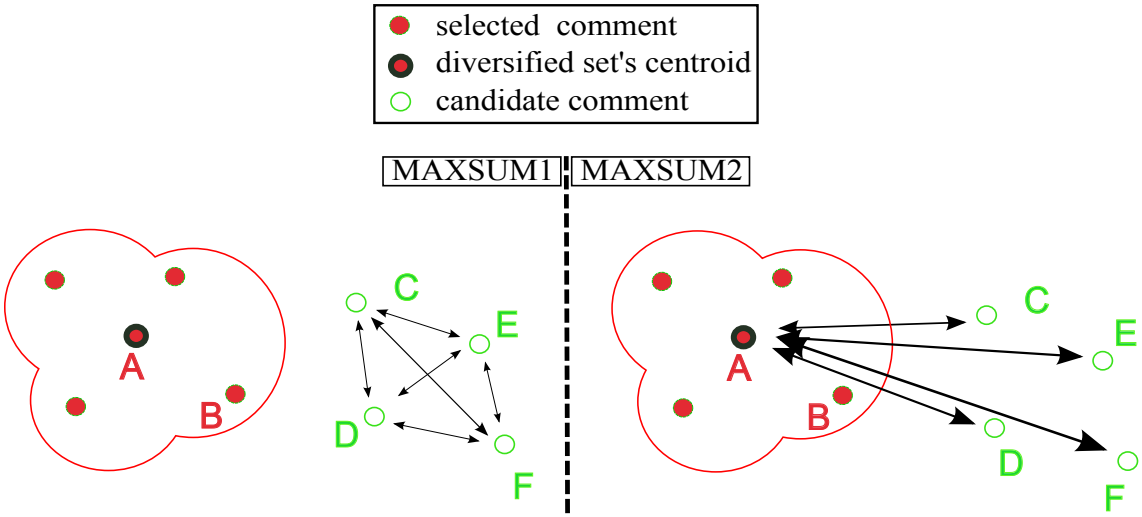
$$d(u, S) = \frac{1}{|S|} \sum_{x \in S} d(u, x) \tag{6.1}$$

Αν και ο αλγόριθμος στοχεύει στη μεγιστοποίηση της ίδιας συνάρτησης στόχου, η κύρια διαφορά του με τον αλγόριθμο 5.2 είναι ότι σε κάθε βήμα εξετάζει τις αποστάσεις μεταξύ υποψηφίων και ήδη επιλεγμένων σχολίων.

Αλγόριθμος 6.1 Αλγόριθμος Διαφοροποίησης Max-sum2

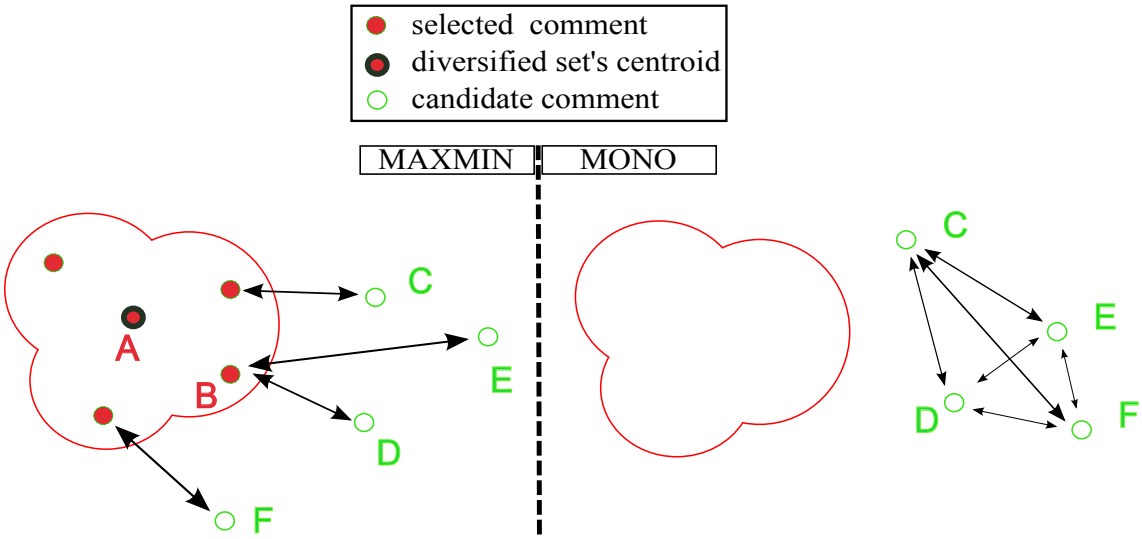
Input: Set of candidate comments N, size of diverse set k
Output: Set of diverse comments $S \subseteq N, |S| = k$
  $S = \emptyset$
  Find the most relevant comment u and set $S = \{u\}$
  For any $x \in N \setminus S$, define $d_{MAX}(x, S) = d(x, c_s)$ ▷ c_s is the centroid of the comments contained in S
  while $|S| < k$ **do**
    FIND $u = argmax_{x \in N} d_{MAX}(x, S)$
    Set $S = S \cup \{u\}$
    Set $N = N \setminus \{u\}$
    Update c_s
  end while

Το Σχήμα 6.3 οπτικοποιεί την λειτουργία των αλγορίθμων διαφοροποίησης Max-sum και Max-sum2. Τα σχόλια που έχουν επιλεγεί στο διαφοροποιημένο σύνολο S αναπαρίσταται ως γεμάτοι κύκλοι, ενώ τα υπόλοιπα υποψήφια σχόλια αναπαρίστανται ως κενοί κύκλοι. Ο αλγόριθμος Max-sum (Αλγόριθμος 5.2) εξετάζει όλα τα ζεύγη αποστάσεων μεταξύ των υποψήφιων σχολίων και επιλέγει τα σχόλια C και F, που έχουν την μεγαλύτερη ανά-μεταξύ τους απόσταση. Ο αλγόριθμος Max-sum2 (Αλγόριθμος 6.1) εξετάζει όλα τα υποψήφια σχόλια με το ιδεατό σχόλιο A, που αναπαριστά το centroid του διαφοροποιημένου συνόλου S, και επιλέγει αυτό με την μεγαλύτερη απόσταση, εν προκειμένω το F. Σημειώνουμε ότι, στο επόμενο βήμα, το centroid του διαφοροποιημένου συνόλου S επανυπολογίζεται, οπότε οι αποστάσεις των υποψήφιων σχολίων από το νέο centroid θα πρέπει να επαναυπολογιστούν επίσης.



Σχήμα 6.3: Οπτικοποίηση Αλγορίθμων Διαφοροποίησης Max-sum και Max-sum2

Το Σχήμα 6.4 οπτικοποιεί την λειτουργία των αλγορίθμων διαφοροποίησης Max-min και Mono-objective. Τα σχόλια που έχουν επιλεγεί στο διαφοροποιημένο σύνολο S αναπαρίστανται ως γεμάτοι κύκλοι, ενώ τα υπόλοιπα υποψήφια σχόλια αναπαρίστανται ως κενοί κύκλοι. Ο αλγόριθμος Max-min (Αλγόριθμος 5.3) υπολογίζει όλες τις ανά δύο αποστάσεις μεταξύ των υποψηφίων και των επιλεγμένων σχολίων. Στη συνέχεια, θα εισάγει το E ως το σχόλιο με την μέγιστη ελάχιστη απόσταση, από τα ήδη επιλεγμένα σχόλια του S. Ο αλγόριθμος Mono-objective (Αλγόριθμος 5.4) θα υπολογίσει, κατά την αρχικοποίηση, ένα σκορ για κάθε υποψήφιο σχόλιο και θα εισάγει σχόλια στο S, βασιζόμενος στα αρχικώς υπολογισμένα σκορ. Στο παράδειγμα μας, το σχόλιο C, που έχει την μέγιστη μέση απόσταση από όλα τα άλλα υποψήφια σχόλια, θα εισαχθεί πρώτο. Παρατηρούμε επίσης ότι ο αλγόριθμος επαναχρησιμοποιεί σε κάθε βήμα τις ήδη υπολογισμένες αποστάσεις των υποψηφίων σχολίων.



Σχήμα 6.4: Οπτικοποίηση Αλγορίθμων Διαφοροποίησης Max-min και Mono-objective

Τα παραπάνω παραδείγματα καταδεικνύουν ότι διαφορετικές μέθοδοι δύναται να επιφέρουν

διαφορετικά τελικά αποτελέσματα, δηλαδή διαφορετικά σύνολα διαφοροποιημένων αποτελεσμάτων και επομένως αναμένουμε σημαντικές διαφορές όσον αφορά την αξιολόγηση της αποτελεσματικότητας κάθε αλγόριθμου.

6.1.5 Συναρτήσεις Αποστάσεων

Στη συνέχεια, περιγράφουμε τις συναρτήσεις αποστάσεων που εφαρμόζονται από τον κάθε αλγόριθμο.

Ανάλογα με τον αλγόριθμο διαφοροποίησης που εφαρμόζεται ορίζουμε τέσσερις τύπους που παράγουν την τελική τιμή βαθμολόγησης για κάθε υποψήφιο σχόλιο u, κειμένου A, το οποίο έχει N σχόλια, προκειμένου να επιλεγεί για εισαγωγή στο διαφοροποιημένο υποσύνολο S, με $\lambda \epsilon [0..1]$ παράμετρος που προσδιορίζει το συμβιβασμό μεταξύ συνάφειας και ανομοιότητας, i είναι το κριτήριο διαφοροποίησης, $|D|$ είναι ο αριθμός των κριτηρίων διαφοροποίησης και $w_i \epsilon [0,1]$ είναι το βάρος του κάθε επιμέρους κριτηρίου, με $\sum_{i=1}^{|D|} w_i = 1$.

– **Max-sum**

$$score_{Max-sum}(u, v, A) = (1 - \lambda) \cdot \frac{r(u, A) + r(v, A)}{2} + \lambda \cdot \sum_{i=1}^{|D|} w_i \cdot d_i(u, v) \qquad (6.2)$$

όπου (u, v) είναι ένα ζεύγος σχολίων, αφού αυτός ο αλγόριθμος εξετάζει τα σχόλια ανά ζεύγη.

– **Max-min**

$$score_{Max-Min}(u, A) = (1 - \lambda) \cdot r(u, A) + \lambda \cdot \sum_{i=1}^{|D|} w_i \cdot d_i(u, \min v_{iu}) \qquad (6.3)$$

όπου $\min v_{iu}$ είναι το σχόλιο από το τρέχον διαφοροποιημένο σύνολο με τη μικρότερη απόσταση από το υποψήφιο σχόλιο u

– **Mono-objective**

$$score_{Mono-objective}(u, A) = (1 - \lambda) \cdot r(u, A) + \lambda \cdot \sum_{i=1}^{|D|} w_i \cdot \frac{1}{|T| - 1} \sum_{v \epsilon T} d(u, v) \qquad (6.4)$$

– **Max-sum2**

$$score_{Max-sum2}(u, A) = (1 - \lambda) \cdot r(u, A) + \lambda \cdot \sum_{i=1}^{4} w_i \cdot d_i(u, C_i) \qquad (6.5)$$

όπου C_i είναι το centroid του τρέχοντος διαφοροποιημένου συνόλου, όσον αφορά τη διάσταση διαφοροποίησης i.

Στην Ενότητα 6.1.3 αναλύθηκαν τα κριτήρια διαφοροποίησης, με βάση τα οποία εκφράζουμε κάθε σχόλιο μέσω διανυσμάτων χαρακτηριστικών, που αντιστοιχούν στα κριτήρια διαφοροποίησης και τα οποία αντιπροσωπεύουν διάφορες εκφάνσεις των σχολίων. Οι αλγόριθμοι διαφοροποίησης (Ενότητα 6.1.4) χρησιμοποιούν αυτά τα διανύσματα για να υπολογίζουν σε κάθε βήμα ένα αθροιστικό σκορ διαφοροποίησης για κάθε σχόλιο. Το τελικό σκορ, που παράγεται σταθμίζοντας το σκορ ομοιότητας του σχολίου με το άρθρο με το σκορ διαφοροποίησης, καθορίζει την επιλογή του επόμενου σχολίου. Το σκορ διαφοροποίησης προκύπτει από την απόσταση μεταξύ των δύο σχολίων. Ως συνάρτηση απόστασης χρησιμοποιούμε την συνάρτηση ομοιότητας συνημίτονου (cosine similarity function), σε κανονικοποιημένη μορφή, και ορίζουμε το σκορ διαφοροποίησης μεταξύ δύο στοιχείων, u, v, με διανυσμάτων χαρακτηριστικών, $V(u), V(v)$, ως προς μία διάσταση-κριτήριο διαφοροποίησης i, ως εξής:

$$d_i(u, v) = 1 - \cos_i(V(u), V(v)) \tag{6.6}$$

Οι τιμές που προκύπτουν από την παραπάνω εξίσωση είναι κανονικοποιημένες στο επίπεδο του κάθε κριτηρίου ξεχωριστά. Δηλαδή, υπολογίζουμε τη μέγιστη τιμή που μπορεί να πάρει κάποιο κριτηρίου για όλα τα σχόλια και διαιρούμε τα αντίστοιχα σκορ των υπολοίπων σχολίων με αυτό, για κάθε κριτήριο ξεχωριστά.

Η ετερογένεια μεταξύ των σχολίων δεν είναι ο μόνος στόχος: τα σχόλια οφείλουν να είναι και σχετικά με το ειδησεογραφικό άρθρο. Συνεπώς το τελικό σκορ για κάθε σχόλιο είναι το σταθμισμένο άθροισμα του συνολικού σκορ διαφοροποίησης του και του σκορ ομοιότητάς του με το άρθρο. Το σκορ ομοιότητας r ενός σχολίου u με ένα άρθρο A, ορίζεται με βάση τη συνάρτηση συνημίτονου στα διανύσματα χαρακτηριστικών όρων του σχολίου και του άρθρου:

$$r(u, A) = \cos(V(u), V(A)) \tag{6.7}$$

Σημειώνουμε ότι και αυτό το σκορ κανονικοποιείται στο διάστημα [0..1].

6.2 Πειραματική Μελέτη

Στην ενότητα αυτή, περιγράφουμε τη συλλογή ειδησεογραφικών άρθρων και σχολίων χρηστών που χρησιμοποιούμε, τα σενάρια αξιολόγησης κριτηρίων και αλγορίθμων που αποτιμήσαμε, την μεθοδολογία που ακολουθήσαμε για την υποσημείωση με κρίσεις συνάφειας των εγγράφων, είτε αυτές ορίζονται από τον χρήστη είτε προκύπτουν αυτόματα με αντικειμενικές μεθόδους, και τις μετρικές που χρησιμοποιούμε για την αξιολόγηση της απόδοσης των μεθόδων. Τέλος, παρέχουμε τα αποτελέσματα μαζί με μια σύντομη συζήτηση.

6.2.1 Συλλογή Ειδησεογραφικών Άρθρων και Σχολίων Χρηστών

Για την αξιολόγηση χρησιμοποιήσαμε ένα σύνολο ειδησεογραφικών άρθρων και των αντιστοίχων σχολίων χρηστών από την εφημερίδα New York Times[2]. Η διαδικτυακή έκδοση

[2] `https://www.nytimes.com/`

της εφημερίδας προσφέρει ένα καλά οργανωμένο API για την ανάκτηση άρθρων και σχολίων Times Developer Network[3]. Κάθε άρθρο / σχόλιο συνοδεύεται από τα μεταδεδομένα του, όπως ημερομηνία, θεματική κατηγοριοποίηση, χρήστης, κτλ. Για να συγκεντρώσουμε μία επαρκή ποσότητα άρθρων, κατεβάσαμε άρθρα μέσω του API, χρησιμοποιώντας τη λέξη-κλειδί "financial". Αυτή η διαδικασία μας επέστρεψε 2800 άρθρα, για καθένα απο τα οποία εξετάσαμε εάν υπάρχει επαρκής αριθμός σχολίων (περισσότερα από 100). Τελικά ανασύραμε 1935 άρθρα με ένα σύνολο 293,303 σχολίων, το οποίο δίνει ένα μέσο όρο 152 σχολίων ανά άρθρο. Σημειώνουμε ότι, αφού (α) η λέξη-κλειδί που χρησιμοποιήσαμε είναι αρκετά γενική και (β) αναζητείται και στο κείμενο των άρθρων, το ανακτηθέν σύνολο άρθρων δεν περιορίζεται μόνο σε οικονομικά άρθρα και είναι αρκετά γενικό ώστε να περιέχει και άλλες θεματικές περιοχές, όπως πολιτική, επιχειρήσεις, κτλ.

Το ευρετήριο μας κατασκευάστηκε χρησιμοποιώντας τυπική λίστα κοινών λέξεων και τεχνική Porter stemming, με log-based $tf - idf$ σχήμα ευρετηρίασης. Για την εξαγωγή των ονοματικών οντοτήτων χρησιμοποιήσαμε το Stanford Named Entity[4] [44] και του συναίσθηματος το Sentistrength [139]. Τέλος, υπολογίστηκαν και κανονικοποιήθηκαν οι τιμές των κριτηρίων διαφοροποίησης, για κάθε άρθρο/ σχόλιο χρήστη της συλλογής δεδομένων μας.

6.2.2 Σενάρια Αξιολόγησης

Σε αυτήν την ενότητα παρουσιάζουμε μία αναλυτική αξιολόγηση των αλγορίθμων και των κριτηρίων που περιγράφηκαν προηγουμένως. Ως βασική μέθοδο προς σύγκριση με τις μεθόδους μας θεωρούμε την απλή, αν και καθιερωμένη στο σενάριο της διαφοροποίησης αποτελεσμάτων αναζήτησης, μέθοδο Content Diversity - CONTENTDIV, η οποία διαφοροποιεί τα σχόλια χρηστών βασιζόμενη μόνο στο κριτήριο της κειμενικής ομοιότητας. Οι υπόλοιπες μέθοδοι αποτελούν αντιπροσωπευτικές παραλλαγές της μεθόδου μας, όσον αφορά στα κριτήρια διαφοροποίησης που συνδυάζουν. Οι συγκρινόμενες μέθοδοι περιγράφονται παρακάτω:

- **Κειμενική διαφοροποιηση - CONTENTDIV**: Η βασική μέθοδος σύγκρισης που εφαρμόζει διαφοροποίηση μόνο πάνω στο κειμενικό περιεχόμενο.
- **Διαφοροποίηση συναισθήματος - SENTIDIV**: Η παραλλαγή που διαφοροποιεί μόνο πάνω στο αναγνωριζόμενο συναίσθημα των σχολίων.
- **Διαφοροποίηση ονοματικών οντοτήτων - NEDIV**: Η παραλλαγή που διαφοροποιεί μόνο πάνω στις αναγνωριζόμενες ονοματικές οντότητες.
- **Διαφοροποίηση συναισθήματος ονοματικών οντοτήτων - NESENTIDIV**: Η παραλλαγή που διαφοροποιεί πάνω στο συναίσθημα που αναγνωρίζεται γύρω από τις αναγνωριζόμενες ονοματικές οντότητες του σχολίου.
- **Υβριδική διαφοροποίηση - SEMIHYBRID**: Η παραλλαγή που διαφοροποιεί συνδυάζοντας τα κριτήρια της κειμενικής ομοιότητας, του συναισθήματος και των ονοματικών οντοτήτων.

[3] `https://developer.nytimes.com/`

[4] `http://nlp.stanford.edu/software/CRF-NER.shtml`

– **Εκτεταμένη υβριδική διαφοροποίηση - HYBRID**: Η παραλλαγή που διαφοροποιεί συνδυάζοντας όλα προτεινόμενα κριτήρια.

Κάθε μία από τις παραπάνω μεθόδους συνδυάστηκε με καθέναν από τους τέσσερις αλγορίθμους διαφοροποίησης που παρουσιάστηκαν στην Ενότητα 6.1.4: Max-sum, Max-sum2, Max-min και Mono-objective.

Σημειώνουμε ότι, στις μεθόδους που συνδυάζουν κριτήρια, τα σκορ των επιμέρους κριτηρίων σταθμίζονται ισοδύναμα ώστε να παραχθεί το τελικό σκορ διαφοροποίησης. Επιπλέον, για όλες τις μεθόδους θέτουμε ένα σταθερό βάρος $\lambda = 0.7$ για το σκορ διαφοροποίησης και αντίστοιχα βάρος $(1 - \lambda) = 0.3$ για το σκορ ομοιότητας με το άρθρο. Με αυτόν τον τρόπο θέλουμε να εξασφαλίσουμε μία ελάχιστη κειμενική ομοιότητα των διαφοροποιημένων σχολίων με το άρθρο, η οποία βοηθά να εξαλειφθούν αποκλίνοντα σχόλια. Καθώς η βαρύτητα της διαφοροποίησης είναι υπερδιπλάσια από τη βαρύτητα της ομοιότητας με το άρθρο, πιστεύουμε ότι το σκορ ομοιότητας δεν επηρεάζει καταλυτικά την επιλογή ετερογενών σχολίων.

6.2.3 Μετρικές Αποτίμησης

Οι μετρικές αποτίμησης που χρησιμοποιούμε βασίζονται στην έννοια των *μονάδων πληροφορίας* (information nuggets). Όπως παρουσιάζονται στο [30], οι μονάδες πληροφορίας πληροφορίας αντιπροσωπεύουν ιδιότητες της πηγής πληροφοριών στο ένα άκρο και συστατικά της πληροφοριακής ανάγκης στο άλλο. Για την εφαρμογή τους στην διαφοροποίηση αποτελεσμάτων αναζήτησης, οι μονάδες μπορεί να είναι διαφορετικές απαντήσεις ενός ερωτήματος, ή διαφορετικές εκφάνσεις θεμάτων /εννοιών που αντιστοιχούν στους όρους του ερωτήματος. Στο δικό μας σενάριο, της διαφοροποίησης σχολίων χρηστών, προσαρμόζουμε τον ορισμό των μονάδων πληροφορίας:

Ορισμός 6.1 (Μονάδα πληροφορίας)**.** *Μονάδα πληροφορίας είναι κάθε έννοια ή θεματική κατηγορία ή υποκατηγορία ή σχετική άποψη / συναίσθημα ή επέκταση των παραπάνω εννοιών και κατηγοριών που σχετίζεται με την καταχώρηση του χρήστη.*

Ο προηγούμενος ορισμός, παρότι δεν είναι αυστηρός, επιτρέπει την αντιστοίχιση της ποικιλομορφίας σε θέματα και έννοιες που εντοπίζονται στο κείμενο των άρθρων και των σχολίων και διευκολύνει των ορισμό κατάλληλων μετρικών διαφοροποίησης:

– **Nugget Coverage - NC@n**: Η μετρική Nugget Coverage ποσοτικοποιεί την κάλυψη των μονάδων πληροφορίας στο σύνολο των αποτελεσμάτων. Πρόκειται για μετρική που βασίζεται στην μετρική Precision @N και υπολογίζει πόσες μονάδες πληροφορίας, συνολικά, εντοπίζονται στις καταχωρήσεις του διαφοροποιημένου συνόλου αποτελεσμάτων. Ορίζεται ως εξής:

$$NC@n = \frac{\sum_{k=1}^{n} I_k}{n \cdot |I|} \tag{6.8}$$

όπου n είναι η θέση στην οποία αποτιμάται, συνήθως 5,10,20, I_k ο αριθμός των διακριτών μονάδων πληροφορίας που περιέχονται στη θέση k και $|I|$ ο συνολικός αριθμός των μονάδων πληροφορίας.

– **Distinct Nugget Coverage - DN@n**: Η μετρική Distinct Nugget Coverage λειτουργεί συμπληρωματικά ως προς την μετρική Nugget Coverage - NC@n και υπολογίζει το λόγο των διακριτών μονάδων πληροφορίας που βρίσκονται στις καταχωρήσεις, προς το συνολικό αριθμό διακριτών μονάδων. Ορίζεται ως εξής:

$$DN@n = \frac{\sum_{i=1}^{|I|} DFI_i}{|I|}, \tag{6.9}$$

με το DFI_i να δίνεται από:

$$DFI_i = \begin{cases} 1 & : FI_i > 0 \\ 0 & : FI_i = 0 \end{cases} \tag{6.10}$$

όπου FI_i είναι η συχνότητα της μονάδας πληροφορίας i στην θέση αποτίμησης n.

– **Nugget Uniformity - NU@n**: Η μετρική Nugget Uniformity ποσοτικοποιεί την διακύμανση των μονάδων πληροφορίας στο σύνολο αποτελεσμάτων, απαιτώντας οι μονάδες να είναι όσο το δυνατόν πιο ομοιόμορφα κατανεμημένες. Υπολογίζει την διακύμανση των συχνοτήτων των μονάδων στο σύνολο αποτελεσμάτων, ως προς την μέση τιμή των συχνοτήτων. Ορίζεται ως εξής:

$$NU@n = \frac{\sum_{i=1}^{|I|}(FI_i - \bar{I})^2}{|I|} \tag{6.11}$$

με την τη μέση τιμή των συχνοτήτων των μονάδων πληροφορίας να δίνεται από:

$$\bar{I} = \frac{\sum_{i=1}^{|I|} FI_i}{|I|} \tag{6.12}$$

6.2.4 Κρίσεις Συνάφειας

Όπως αναφέρθηκε και στην Ενότητα 5.3.3, μια από τις δυσκολίες στην αξιολόγηση μεθόδων που έχουν σχεδιαστεί για την διαφοροποίηση αποτελεσμάτων αποτελεί η έλλειψη τυπικών δεδομένων δοκιμών, η έλλειψη ενός συνόλου αντικειμενικής αλήθειας. Στην περίπτωση μας, διαθέτοντας μόνο την συλλογή άρθρων/σχολίων χρηστών, χρειάστηκε να καθορίσουμε επιπρόσθετα: (α) μια μέθοδο για τον προσδιορισμό των μονάδων πληροφορίας σε κάθε άρθρο και (β) μια μέθοδο για την υποσημείωση των σχολίων χρηστών με κρίσεις συνάφειας για το κάθε θέμα. Με βάση την έλλειψη τυπικών δοκιμαστικών δεδομένων, για να εκτιμήσουμε και να αξιολογήσουμε τις επιδόσεις των διαφόρων μεθόδων διαφοροποίησης στην συλλογή καταχωρήσεων μας, μεριμνήσαμε για την δημιουργία του συνόλου αντικειμενικής αλήθειας τόσο από τους χρήστες, όσο και αυτόματα από αντικειμενικές μεθόδους.

Κρίσεις Συνάφειας οριζόμενες από τον Χρήστη

Επιλέξαμε τυχαία 10 άρθρα και το σύνολο των σχολίων για το καθένα. Για κάθε ένα από τα σενάρια αξιολόγησης (Ενότητα 6.2.2), με εφαρμογή σε κάθε έναν από τους αλγορίθμους, επιστρέφουμε ένα σύνολο από top-10 διαφοροποιημένα σχόλια. Δεδομένου ότι κάθε σενάριο εφαρμόζεται σε κάθε έναν από τους αλγορίθμους, καταλήγουμε σε 24 συνδυασμούς προς αξιολόγηση.

Κρατάμε μόνο τα αναγνωριστικά των σχολίων και αφαιρούμε κάθε πληροφορία προέλευσης του σχολίου, δηλαδή από ποιες από τις 24 μεθόδους προέκυψε. Στη συνέχεια, με την βοήθεια δύο εξωτερικών συνεργατών, εκτελούμε τα ακόλουθα δύο βήματα:

- **α) Εξαγωγή Μονάδων Πληροφορίας από τον Χρήστη**. Διαβάσαμε κάθε άρθρο και τα σχόλια του, από τους 24 συνδυασμούς προς αξιολόγηση, και δημιουργήσαμε ένα σύνολο με όλες τις διακριτές μονάδες πληροφορίας που σχετίζονται με τα θέματα του άρθρου. Με αυτόν τον τρόπο φτιάξαμε μία δεξαμενή από μονάδες πληροφορίας. Επιλέξαμε να εξάγουμε μονάδες πληροφορίας, εκτός από το άρθρο, και από τα σχόλιά του, αφού, θεωρούμε τα σχόλια πολύτιμα μεταδεδομένα του άρθρου, που το συμπληρώνουν και το εμπλουτίζουν. Σημειώνουμε, επίσης, ότι αυτή η διαδικασία περιελάμβανε περισσότερες από μία επαναλήψεις ανά ομάδα άρθρου-σχολίων: ορισμένες φορές, αφού ανακαλύπταμε μία νέα μονάδα πληροφορίας, επιστρέφαμε πίσω και επανεξετάζαμε το άρθρο και τα προηγούμενα σχόλια, έτσι ώστε να βεβαιωθούμε ότι σε κάθε σχόλιο θα είχε ανατεθεί κάθε σχετική με αυτό μονάδα πληροφορίας.

- **β)Επισημείωση σχολίων**. Στη συνέχεια, αναθέσαμε στους δύο εξωτερικούς κριτές την επισημείωση των σχολίων με μονάδες πληροφορίας. Στους κριτές δόθηκαν: (α) το κείμενο έκαστου άρθρου, (β) το διακριτό σύνολο σχολίων που προκύπτουν από τα πρώτα 10 σχόλια, από όλες τις μεθόδους, απαλλαγμένο από πληροφορία προέλευσης, ώστε να μην επηρεαστούν, για παράδειγμα, παρατηρώντας κάποιο μοτίβο προέλευσης σχολίων από συγκεκριμένη μέθοδο και (γ) το σύνολο των μονάδων πληροφορίας. Από τους κριτές ζητήθηκε να επισημειώσουν κάθε σχόλιο με τις μονάδες πληροφορίας που κρίνουν ότι περιέχει. Σημειώνουμε ότι δεν επιβαρύναμε τους επισημειωτές με τη διαδικασία της αναγνώρισης μονάδων πληροφορίας, καθώς μπορούσαν να επιλέξουν μόνο μονάδες από τη δεξαμενή του προηγούμενου βήματος (α). Ταυτόχρονα ο διαχωρισμός μεταξύ χρηστών εξαγωγέων πληροφορίας και χρηστών επισημείωσης σχολίων εξασφαλίζει την αντικειμενικότητα της διαδικασίας.

Ουσιαστικά, οι μονάδες πληροφορίες, είναι μια βοηθητική έννοια που ορίζουμε προκειμένου να ποσοτικοποιηθεί η ποικιλομορφία/διαφορετικότητα. Στόχος είναι η διάσπαση της γενικής έννοια της διαφοροποίησης σε δομικές/ελάχιστες μονάδες πληροφοριών, όπου κάθε μια εξ' αυτών αντιπροσωπεύει μια διαφορετική πτυχή των θεμάτων του άρθρου. Με αυτόν τον τρόπο η διαφοροποίηση μπορεί να προσδιοριστεί ποσοτικά και εν συνεχεία να αξιολογηθεί. Επομένως αντί να ζητήσουμε από τον χρήστη να σχολιάσει σε γενικές γραμμές τη διαφορετικότητα/ποικιλομορφία ενός συνόλου σχολίων, του ζητάμε να επισημειώσει τα σχόλια με

Article's main topic	**Indicative nuggets**
Tax evasion	Tax evasion, Ethics, Law and Legislation, Politics
Obama's anti-foreclosure plan	Housing bubble, Politics, People's irresponsibility, Mortgage crisis
Scandal with politician and bankers	Bailout, Bank's name, Legislation, Corruption, Discrimination
Financial reform related to elections	Goldman Sachs, Subprime mortgage crisis, Economic ideologies, Wall Street
Federal loans on energy programs	Alternative energy, Politics, Competitiveness, Financial crisis, Political parties
US consumption rate	Consumption comparisons, Free-market economy, Solutions, Self criticism
Democrats nominations	Clintons unite Republicans, Obama represents change, Critisism on candidates
Prescriptions decrease: consequences / reasons	Critisism on corporations, Economical drug solutions, Patients examples
Obama measures on financial crisis	Critisize irresponsible americans, Blame free market, Measures are moderate
Relation between successful people / elite colleges	Community college inferior to others, How students exploit education matters

Πίνακας 6.1: Χρησιμοποιούμενα στην αξιολόγηση άρθρα, μαζί με, τυχαία επιλεγμένες, ενδεικτικές μονάδες πληροφορίας για το καθένα, όπως εξάχθηκαν από τους αξιολογητές.

ένα σύνολο ήδη εξαχθέντων μονάδων πληροφορίας. Αυτή η διαδικασία μειώνει την πολυπλοκότητα του έργου σχολιασμού και καθιστά την αξιολόγηση ανεξάρτητη των κριτηρίων που χρησιμοποιούνται καθώς ενώ ο υπολογισμός της διαφορετικότητας χρησιμοποιεί μεταξύ άλλων το κειμενικό περιεχόμενο των άρθρων/ σχολίων, οι μονάδες πληροφορίες ορίζονται από τον χρήστη και μπορεί να μην περιέχονται αυτούσιες, να μην αντιστοιχούν σε ακριβείς λέξεις ή φράσεις από το κείμενο των σχολίων.

Στον Πίνακα 6.1 παρουσιάζουμε τα δέκα χρησιμοποιούμενα στην αξιολόγηση άρθρα, μαζί με, τυχαία επιλεγμένες, ενδεικτικές μονάδες πληροφορίας για το καθένα, όπως εξάχθηκαν από την προαναφερθείσα διαδικασία.

Κρίσεις Συνάφειας οριζόμενες αυτόματα

Οι μονάδες πληροφορίας για τα άρθρα και τα σχόλια τους είναι δυνατόν να εξαχθούν και με αυτόματο τρόπο. Στην κατεύθυνση αυτή χρησιμοποιήσαμε τις υπηρεσίες OpenCalais Web Service[5] και AlchemyAPI[6], οι οποίες επιστρέφουν σημασιολογικά μεταδεδομένα για κάθε τεθέν κείμενο. Υποβάλαμε κάθε ένα από τα άρθρα, και εν συνεχεία κάθε σχόλιο εκάστου άρθρου στις εν λόγω υπηρεσίες και σχηματίσαμε ένα σύνολο μεταδεδομένων τόσο σε επίπεδο άρθρου, όσο και σχολίου. Τα μεταδεδομένα αυτά εν συνεχεία τα θεωρούμε ως μονάδες πληροφορίας τόσο για κάθε άρθρο, όσο και κάθε σχόλιο. Με τον τρόπο αυτό, στηριζόμενοι σε

[5] `http://www.opencalais.com/`
[6] `http://www.alchemyapi.com/api/concept/`

Article's main topic	**Indicative nuggets**
Tax evasion	Taxation in the United States, Barack Obama, Moonlight, Banking in Switzerland
Obama's anti-foreclosure plan	United States housing bubble, Payments, Labor, Hudson River, Price, Receipt
Scandal with politician and bankers	Law_Crime, California, Subprime mortgage crisis, Military personnel, Question
Financial reform related to elections	Subprime mortgage crisis, Illinois, Long-Term Capital Management, Good, Vaccination
US consumption rate	Political repression, Environmental issues, Broadsheet, Wheat, Question
Democrats nominations	North Carolina, Joe Biden, Southern hip hop, Criticism, President of the United States
Prescriptions decrease: consequences / reasons	Drug rehabilitation, Americas, Stroke, Public choice theory, Moon
Obama measures on financial crisis	Economics terminology, Taxation in the United States, Newspaper, Real estate bubble
Relation between successful people / elite colleges	Student financial aid, SAT, Emotion, Committee on Institutional Cooperation

Πίνακας 6.2: Χρησιμοποιούμενα στην αξιολόγηση άρθρα, μαζί με, τυχαία επιλεγμένες, ενδεικτικές μονάδες πληροφορίας για το καθένα, όπως εξάχθηκαν με αντικειμενικό τρόπο.

αντικειμενικές μεθόδους, επισημειώσαμε την συλλογή καταχωρήσεων μας με κρίσεις μονάδων πληροφορίας.

Στον Πίνακα 6.2, παρουσιάζουμε τα δέκα χρησιμοποιούμενα στην αξιολόγηση άρθρα, μαζί με, τυχαία επιλεγμένες, ενδεικτικές μονάδες πληροφορίας για το καθένα, όπως εξάχθηκαν με αντικειμενικό τρόπο. Παρατηρούμε ότι υπάρχουν ομοιότητες σε μονάδες πληροφορίας, όπως αυτές προέκυψαν με τον αντικειμενικό τρόπο, σε σχέση με την εξαγωγή τους από τον χρήστη. Επιπρόσθετα παρατηρούμε ότι οι αυτόματα εξαγόμενες μονάδες πληροφορίας, ταυτίζονται πολλές φορές με όρους και φράσεις που απαντώνται στο κείμενο των σχολίων. Με τον τρόπο αυτό πολλές από αυτές ενδεχομένως να εμφανίζουν μικρή συσχέτιση με το περιεχόμενο του άρθρου ή και να εξειδικεύουν σε μεγάλο βαθμό μια έννοια, η οποία δεν δύναται να προσδιοριστεί σε άλλα σχόλια.

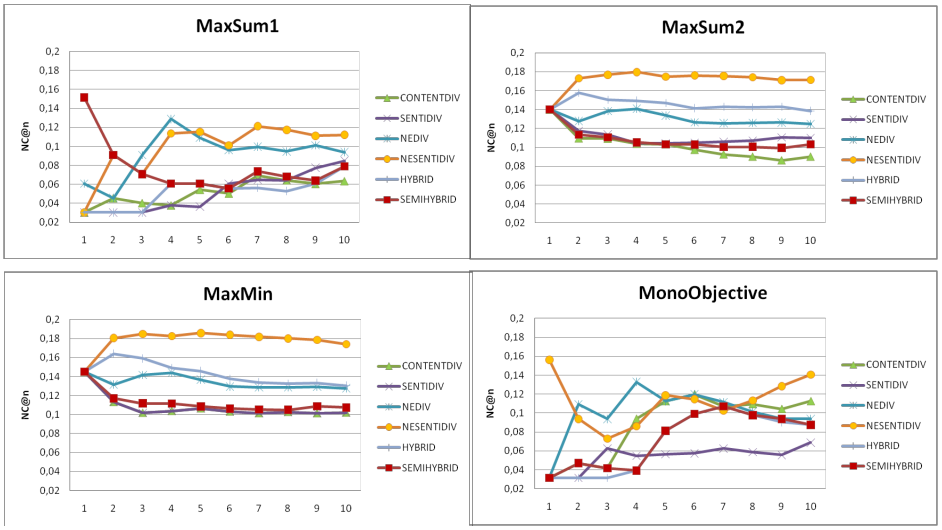
6.2.5 Αποτελέσματα

Στη συνέχεια, παρουσιάζουμε τα αποτελέσματα με βάση τα ανεξαρτήτως παραχθέντα σύνολα αντικειμενικής αλήθειας που κατασκευάσαμε.

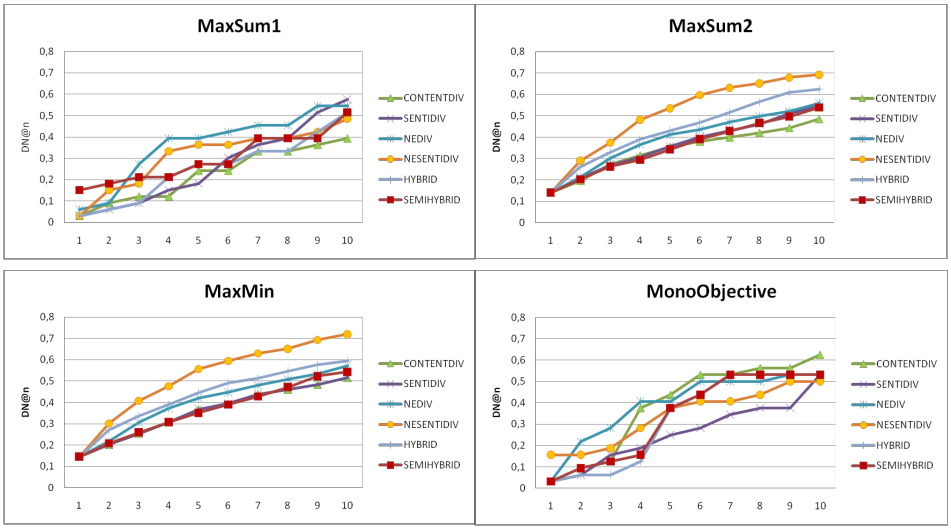
Αποτελέσματα Αξιολόγησης Χρηστών

Στα Σχήματα 6.5 και 6.6 παρουσιάζουμε τις γραφικές παραστάσεις των Nugget Coverage και Distinct Nugget Coverage αντίστοιχα, για τις θέσεις κατάταξης σχολίων από 1 έως 10. Οι γραφικές παρουσιάζονται ανά αλγόριθμο, για ευκολία παρουσίασης. Υπενθυμίζουμε ότι και οι

δύο μετρικές είναι κανονικοποιημένες στο διάστημα [0,1]. Για την μετρική Nugget Coverage, παρατηρούμε ότι οι αλγόριθμοι Max-sum2 και Max-min είναι διακριτά πιο αποτελεσματικοί από τους υπόλοιπους και ταυτόχρονα ακολουθούν μία πιο ομαλή συμπεριφορά, όσον αφορά τις θέσεις κατάταξης, για όλες τις αξιολογούμενες παραλλαγές. Η δεύτερη παρατήρηση είναι ότι η βασική μέθοδος σύγκρισης (CONTENTDIV) αποδίδει σχεδόν πάντα χειρότερα από τις προτεινόμενες παραλλαγές, ανεξαρτήτως αλγορίθμου. Τέλος, η παραλλαγή που συνδυάζει ονοματικές οντότητες και συναίσθημα (NESENTIDIV), είναι διακριτά αποτελεσματικότερη από όλες τις άλλες παραλλαγές και τη βασική μέθοδο, ακολουθούμενη από την παραλλαγή που συνδυάζει όλα τα κριτήρια (HYBRID) και την παραλλαγή των ονοματικών οντοτήτων (NEDIV).



Σχήμα 6.5: Nugget Coverage ανά Αλγόριθμο. (Καλύτερη απεικόνιση έγχρωμα.)



Σχήμα 6.6: Distinct Nugget Coverage ανά Αλγόριθμο. (Καλύτερη απεικόνιση έγχρωμα.)

Για τη μετρική Distinct Nugget Coverage ισχύουν οι ίδιες παρατηρήσεις. Εξαίρεση απο-

τελεί η παραλλαγή CONTENTDIV η οποία εμφανίζεται αποτελεσματικότερη στον αλγόριθμο Mono-objective, αλλά έχει χειρότερη απόδοση σε σχέση με την παραλλαγή NESENTIDIV στους αλγορίθμους Max-sum2 και Max-min.

Οι τιμές των μετρικών Nugget Coverage και Distinct Nugget Coverage, μόνο για τις 5 και 10 πρώτες θέσεις κατάταξης, παρουσιάζονται, ξεχωριστά, στους Πίνακες 6.3, 6.4, 6.5 και 6.6, προκειμένου να τονιστούν καλύτερα οι διαφορές μεταξύ των μεθόδων. Ταυτόχρονα, στους πίνακες αυτούς, σημειώνουμε τον καλύτερο αλγόριθμο ανά παραλλαγή κριτηρίων (τελευταία γραμμή), την καλύτερη παραλλαγή κριτηρίων ανά αλγόριθμο, (τελευταία στήλη) και τον συνολικά καλύτερο συνδυασμό (παραλλαγή/αλγόριθμος). Ο συνδυασμός Max-min/NESENTIDIV ξεπερνάει όλους τους άλλους συνδυασμούς σε αποτελεσματικότητα. Επιπλέον, βελτιώνει την αποτελεσματικότητα της βασικής μεθόδου κατά 65% 54% 27% και 15% για τα επίπεδα NC@5, NC@10, DN@5 και DN@10 αντίστοιχα. Οι επί τοις εκατό διαφορές μεταξύ του Max-min/ NESENTIDIV και της βασικής μεθόδου είναι αντίστοιχα: 7.3% 6.1% 11.8% και 9.5%.

Επίσης, αναφέρουμε τιμές στατιστικής σημασίας των διαφορών μεταξύ των παραλλαγών και της βασικής μεθόδου, χρησιμοποιώντας το t-test με διάστημα εμπιστοσύνης 95%. Σε κάθε γραμμή των Πινάκων 6.3, 6.4, 6.5 και 6.6, σημειώνονται με $*$ οι παραλλαγές εκείνες των οποίων η διαφορά στην αποτελεσματικότητα από τη βασική μέθοδο CONTENTDIV είναι στατιστικά σημαντική. Σημειώνουμε ότι ο καλύτερος συνδυασμός παραλλαγής/αλγορίθμου, Max-min/NESENTIDIV είναι στατιστικά σημαντικά καλύτερος από οποιονδήποτε συνδυασμό βασικής μεθόδου/αλγορίθμου.

Algorithm	**CONTENTDIV**	**SENTIDIV**	**NEDIV**	**NESENTIDIV**	**HYBRID**	**SEMIHYBRID**	**Best Criterion per Algorithm**
Max-sum	0.055	0.036	0.109*	0.115*	0.061*	0.061	NESENTIDIV
Max-sum2	0.103	0.104	0.134	0.175*	0.147*	0.103	NESENTIDIV
Max-min	0.106	0.123	0.137	**0.186***	0.146	0.108	NESENTIDIV
Mono-objective	0.113	0.056	0.113	0.119*	0.081	0.081	NESENTIDIV
Best Algorithm per criterion	Mono-objective	Max-min	Max-min	Max-min	Max-sum2	Max-min	Max-min/ NESENTIDIV

Πίνακας 6.3: Nugget Coverage στην θέση 5

Algorithm	**CONTENTDIV**	**SENTIDIV**	**NEDIV**	**NESENTIDIV**	**HYBRID**	**SEMIHYBRID**	**Best Criterion per Algorithm**
Max-sum	0.064	0.085	0.094	0.112*	0.079*	0.079	NESENTIDIV
Max-sum2	0.090	0.110	0.125	0.171*	0.139*	0.103	NESENTIDIV
Max-min	0.102	0.115	0.128	**0.174***	0.131	0.107	NESENTIDIV
Mono-objective	0.113	0.069	0.094	0.141*	0.088	0.088	NESENTIDIV
Best Algorithm per criterion	Mono-objective	Max-min	Max-min	Max-min	Max-sum2	Max-min	Max-min/ NESENTIDIV

Πίνακας 6.4: Nugget Coverage στην θέση 10

Εξετάζοντας τις παραπάνω γραφικές και πίνακες προκύπτει ότι οι ονοματικές οντότητες είναι ένα σημαντικό κριτήριο διαφοροποίησης των σχολίων χρηστών. Η αποτελεσματικότητα του κριτηρίου ενισχύεται ακόμα περισσότερο, όταν συνδυάζεται με το κριτήριο αναγνώρισης συναισθήματος γύρω από τις ονοματικές οντότητες. Αποδίδουμε αυτήν την συμπεριφορά στο γεγονός ότι τα θέματα που περιγράφονται σε ειδησεογραφικά άρθρα πιθανότατα σχετίζονται με

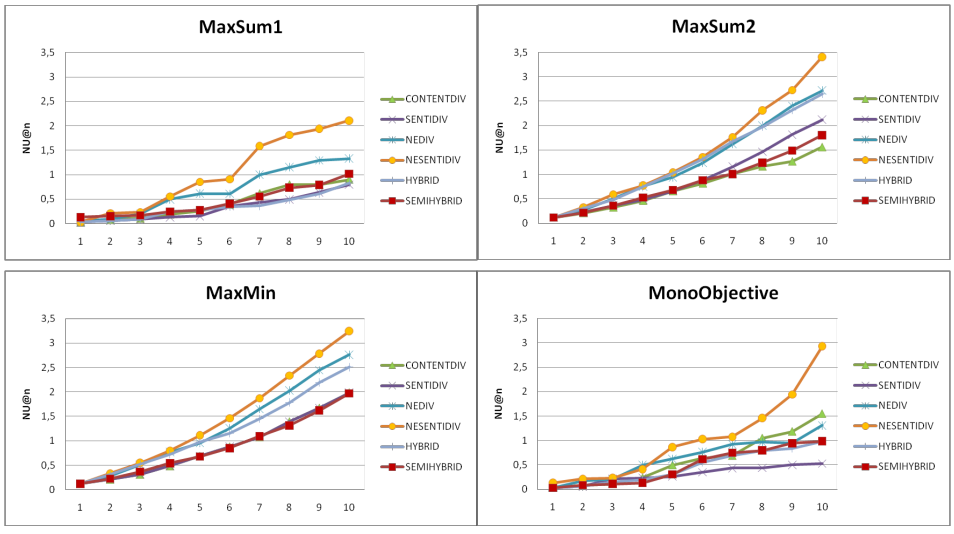
Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
Max-sum	0.242	0.182	0.394	0.364*	0.273*	0.273	NEDIV
Max-sum2	0.355	0.357	0.411	0.536*	0.431	0.342	NESENTIDIV
Max-min	0.368	0.399	0.421	**0.556***	0.445	0.351	NESENTIDIV
Mono-objective	0.438	0.250	0.406	0.375	0.375	0.375	CONTENTDIV
Best Algorithm per criterion	Mono-objective	Max-min	Max-min	Max-min	Max-min	Mono-objective	Max-min/ NESENTIDIV

Πίνακας 6.5: Distinct Nugget Coverage στην θέση 5

Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
Max-sum	0.394	0.576	0.546	0.485*	0.515*	0.515	SENTIDIV
Max-sum2	0.485	0.545	0.560	0.692*	0.625*	0.538	NESENTIDIV
Max-min	0.516	0.557	0.572	**0.720***	0.594	0.543	NESENTIDIV
Mono-objective	0.625	0.531	0.531	0.500*	0.531	0.531	CONTENTDIV
Best Algorithm per criterion	Max-min	Max-min	Max-min	Max-min	Max-sum2	Max-min	Max-min/ NESENTIDIV

Πίνακας 6.6: Distinct Nugget Coverage στην θέση 10

ονοματικές οντότητες, οπότε αυτές βοηθούν στην αποτελεσματικότερη ανίχνευση αυτών των θεμάτων. Ταυτόχρονα, η ετερογένεια συναισθήματος στις ονοματικές οντότητες πιθανότατα υποδηλώνει και ετερογένεια των αντιστοίχων θεμάτων. Παράλληλα, καταδεικνύεται ότι η χρήση πιο εξειδικευμένων κριτηρίων από την απλή κειμενική ομοιότητα επιφέρει αξιοσημείωτα διαφοροποιημένα υποσύνολα σχολίων χρηστών. Τέλος, ο αλγόριθμος Max-min είναι ο πιο αποτελεσματικός, ακολουθούμενος από τον Max-sum2. Και οι δύο αποδίδουν διακριτά καλύτερα από τους Max-sum και Mono-objective. Η κύρια διαφορά μεταξύ των δύο καλύτερων και των δύο χειρότερων αλγορίθμων είναι ότι οι πρώτοι υπολογίζουν σε κάθε βήμα, αποστάσεις μεταξύ υποψήφιων και ήδη επιλεγμένων σχολίων, ενώ οι δεύτεροι, αποστάσεις μόνο μεταξύ υποψηφίων σχολίων.



Σχήμα 6.7: Nugget Uniformity ανά Αλγόριθμο. (Καλύτερη απεικόνιση έγχρωμα.)

Το Σχήμα 6.7 και οι Πίνακες 6.7 και 6.8 παρουσιάζουν τα αποτελέσματα για την μετρική Nugget Uniformity. Η μετρική αυτή, ποσοτικοποιεί τις διαφορές στις συχνότητες των

μονάδων πληροφορίας σε κάθε διαφοροποιημένο αποτέλεσμα και δεν είναι κανονικοποιημένη. Χαμηλότερες τιμές υποδηλώνουν μεγαλύτερη αποτελεσματικότητα. Σε αυτήν την μετρική, αν και η συνολικά καλύτερη απόδοση για τις μετρικές NU@5, NU@10 επιτυγχάνεται από την παραλλαγή SENTIDIV, δεν μπορούμε να ξεχωρίσουμε εάν υπάρχει κάποια παραλλαγή η οποία να ξεπερνά συστηματικά όλες τις υπόλοιπες. Ταυτόχρονα, παρατηρούμε μια αναστροφή των προηγούμενων τάσεων: οι αλγόριθμοι/παραλλαγές που αποδίδουν καλά στις προηγούμενες μετρικές, υπό-αποδίδουν σε αυτήν την μετρική, και το αντίθετο. Αποδίδουμε το εύρημα αυτό στο ότι με την αύξηση του αριθμού των μονάδων πληροφορίας σε ένα διαφοροποιημένο σύνολο σχολίων, οι πιο δημοφιλείς μονάδες συνεισφέρουν περισσότερο στην αύξηση, ενώ όταν ο αριθμός των συνολικών μονάδων είναι μικρός, οι διαφορές στις επιμέρους συχνότητες των διαφορετικών μονάδων αναμένονται επίσης μικρές. Παράλληλα πιστεύουμε ότι υπάρχουν αποκλίνοντες μονάδες πληροφορίας που απαντώνται σε ελάχιστα σχόλια. Όταν ο συνολικός αριθμός μονάδων αυξάνεται, οι συγκεκριμένες μονάδες αναμένεται να παραμένουν λίγες, επηρεάζοντας τις τιμές της μετρικής Nugget Uniformity. Πιστεύουμε ότι οι μετρικές Nugget Coverage και Distinct Nugget Coverage αποτυπώνουν καλύτερα την απόδοση των μεθόδων διαφοροποίησης,και κατα συνέπεια η βέλτιστη μέθοδος διαφοροποίησης θα έπρεπε να επιλέγεται κυρίως με βάση αυτές. Από την άλλη πλευρά, υπάρχουν συνδυασμοί μέσης λύσης, π.χ., οι παραλλαγές NEDIV, HYBRID για τους αλγορίθμους Max-min και Max-sum2. Σημειώνουμε επίσης ότι αντίθετα με τις δύο πρώτες μετρικές, μετρικές ακρίβειας, δε θεωρούμε χρήσιμο να πραγματοποιήσουμε δοκιμές στατιστικής σημαντικότητας στη μετρική Nugget Uniformity, καθώς η μετρική αυτή δεν είναι ούτε κανονικοποιημένη, αλλά ούτε και τα αποτελέσματα μιας τέτοιας δοκιμής θα είχαν διαισθητική σημασία.

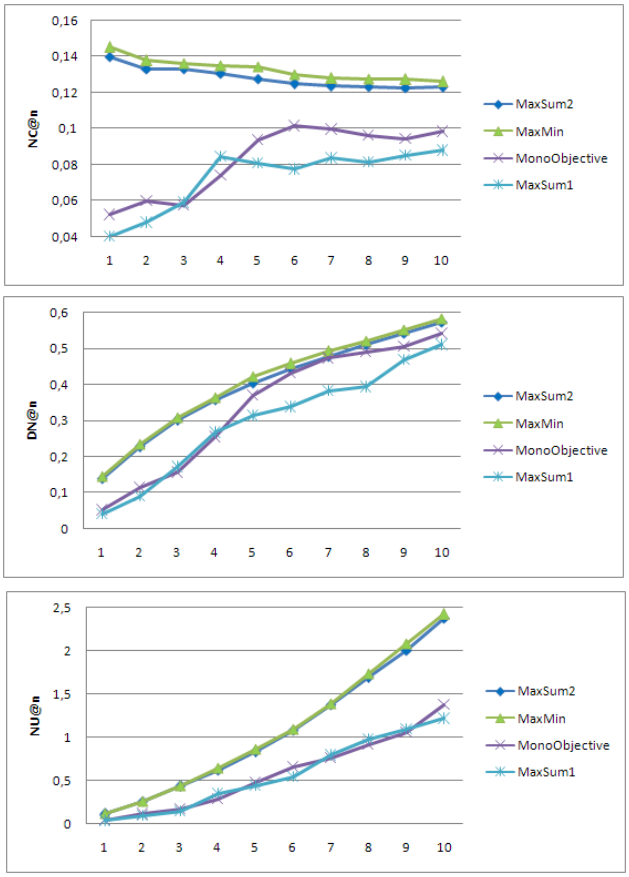
Algorithm	**CONTENTDIV**	**SENTIDIV**	**NEDIV**	**NESENTIDIV**	**HYBRID**	**SEMIHYBRID**	**Best Criterion per Algorithm**
Max-sum	0.259	**0.149**	0.612	0.850	0.272	0.272	SENTIDIV
Max-sum2	0.659	0.662	0.940	1.041	1.018	0.678	CONTENTDIV
Max-min	0.688	0.793	0.954	1.112	0.974	0.677	SEMIHYBRID
Mono-objective	0.496	0.265	0.621	0.866	0.304	0.304	SENTIDIV
Best Algorithm per criterion	Max-sum	Max-sum	Max-sum	Max-sum	Max-sum	Max-sum	Max-sum/ SENTIDIV

Πίνακας 6.7: Nugget Uniformity στην θέση 5

Algorithm	**CONTENTDIV**	**SENTIDIV**	**NEDIV**	**NESENTIDIV**	**HYBRID**	**SEMIHYBRID**	**Best Criterion per Algorithm**
Max-sum	0.898	0.795	1.330	2.107	0.834	1.016	SENTIDIV
Max-sum2	1.565	2.122	2.718	3.407	2.649	1.805	CONTENTDIV
Max-min	1.983	2.079	2.760	3.245	2.509	1.969	SEMIHYBRID
Mono-objective	1.547	**0.5273**	1.309	2.929	0.984	0.984	SENTIDIV
Best Algorithm per criterion	Max-sum	Mono-objective	Mono-objective	Max-sum	Max-sum	Mono-objective	Mono-objective/ SENTIDIV

Πίνακας 6.8: Nugget Uniformity στην θέση 10

Στο Σχήμα 6.8, παρουσιάζεται η μέση τιμή της αποτελεσματικότητας κάθε αλγορίθμου πάνω σε όλες τις παραλλαγές. Οι γραφικές επιβεβαιώνουν τα προηγούμενα αποτελέσματα, όσον αφορά την υπεροχή των αλγορίθμων Max-min και Max-sum2 στις μετρικές κάλυψης μονάδων και την υπό-απόδοση τους στην μετρική ομοιομορφίας.



Σχήμα 6.8: Μέση αποτελεσματικότητα κάθε αλγορίθμου για όλες τις παραλλαγές.(Καλύτερη απεικόνιση έγχρωμα.)

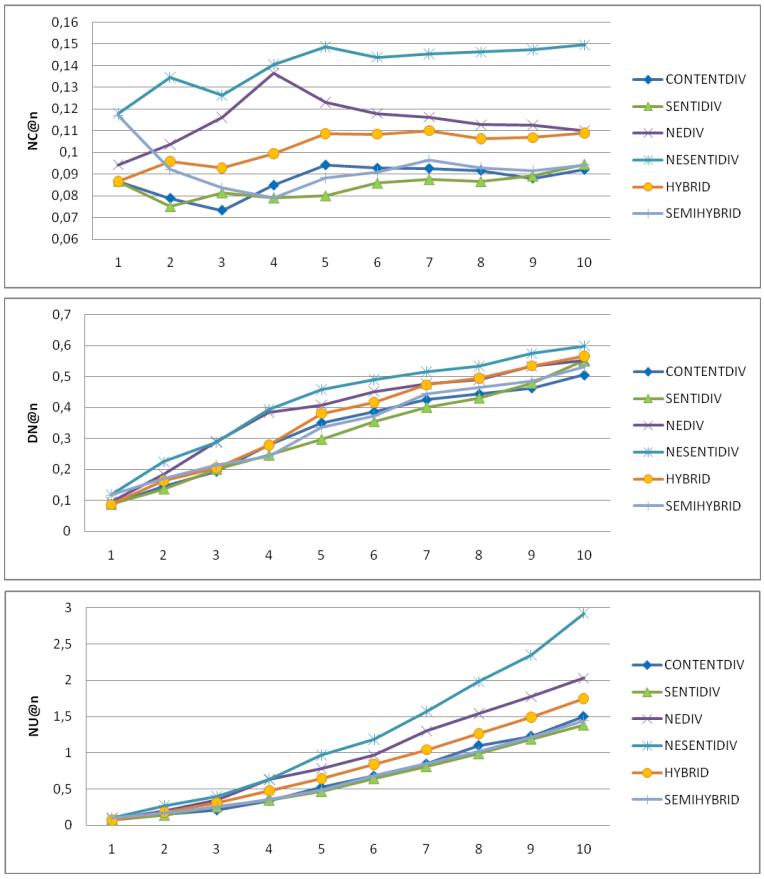
Τέλος, στο Σχήμα 6.9, παρουσιάζεται η μέση αποτελεσματικότητα κάθε παραλλαγής σε όλους τους αλγορίθμους. Τα αποτελέσματα επαληθεύουν τις προηγούμενες παρατηρήσεις σχετικά με την στατιστικά σημαντική υπεροχή της παραλλαγής NESENTIDIV στις μετρικές κάλυψης μονάδων καθώς και της θεώρησης των παραλλαγών NEDIV και HYBRID ως συμβιβαστικές μέσες λύσεις.

Αποτελέσματα Αυτομάτης Αξιολόγησης

Για λόγους απλότητας, η παρουσίαση μας περιλαμβάνει μόνο την μέση αποτελεσματικότητα κάθε αλγόριθμου για όλες τις διαφορετικές παραλλαγές καθώς και την μέση αποτελεσματικότητα κάθε παραλλαγής για όλους τους αλγορίθμους, όπως αποτυπώνονται στα Σχήματα 6.10 και 6.11 αντίστοιχα.

Σε σχέση με τα προηγούμενα αποτελέσματα, παρατηρούμε ότι όσον αφορά την μέση απόδοση του κάθε αλγορίθμου σε όλα τα κριτήρια, Σχήμα 6.10, οι αλγόριθμοι *Max-sum* και *Mono-objective* ξεπερνούν τους *Max-min* και *Max-sum2* στις μετρικές Nugget Coverage και Distinct Nugget Coverage, ενώ η απόδοση όλων των αλγορίθμων στην μετρική Nugget Uniformity είναι παρόμοια.

Επίσης, η μέση αποτελεσματικότητα κάθε παραλλαγής για όλους τους αλγορίθμους εμ-

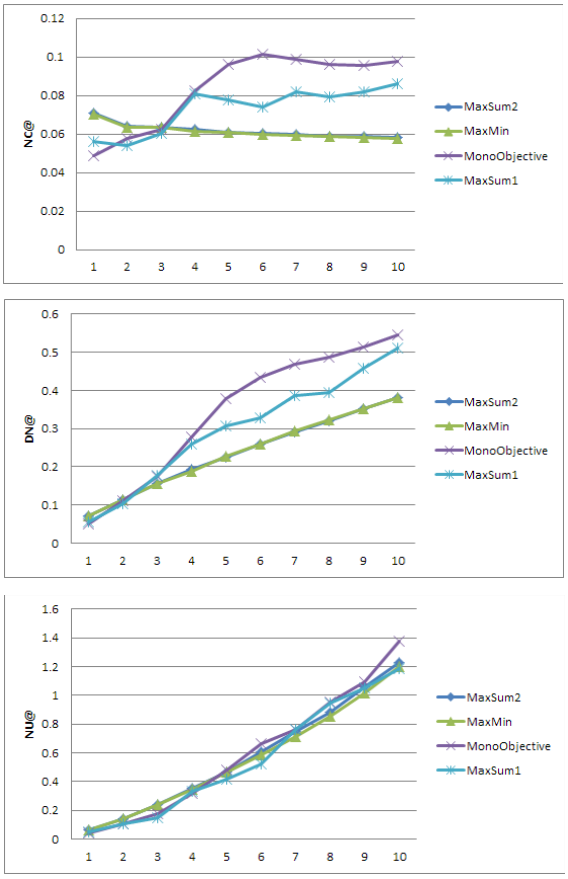


Σχήμα 6.9: Μέση αποτελεσματικότητα κάθε παραλλαγής για όλους τους αλγορίθμους. (Καλύτερη απεικόνιση έγχρωμα.)

φανίζει την ίδια εικόνα με τα προηγούμενα αποτελέσματα, αυτά της αξιολόγησης από τους χρήστες. Αναλυτικότερα η παραλλαγή NESENTIDIV εξακολουθεί να αποδίδει καλύτερα όσον αφορά την κάλυψη, ωστόσο η παραλλαγή NEDIV υπολείπεται ελάχιστα, ενώ αποδίδει καλύτερα στην μετρική Nugget Uniformity. Επιπρόσθετα, για την μετρική Distinct Nugget Coverage, όλα τα κριτήρια φαίνεται να συγκλίνουν, στην θέση 10 ($DN@10$) σε ένα στενό εύρος τιμών, με τις παραλλαγές NEDIV και NESENTIDIV να είναι εμφανώς καλύτερες. Το εύρημα αυτό αποδίδεται στην υπερ-εξειδίκευση των μονάδων πληροφοριών που εισάγουν οι υπηρεσίες αυτόματης εξαγωγής τους, σε σχέση με την ανθρώπινη κρίση, καθώς η τελευταία βασίζεται περισσότερο σε έννοιες και όχι σε όρους που απαντώνται στο κείμενο.

Σύνθεση αποτελεσμάτων

Με βάση τα αποτελέσματα από την αξιολόγηση που πραγματοποιήσαμε, στηριζόμενοι σε κρίσεις συνάφειας των εγγράφων, που προέκυψαν αυτόματα με αντικειμενικές μεθόδους ή ορίστηκαν από τον χρήστη συμπεραίνουμε ότι: (α) Η αποτελεσματικότητα των κριτηρίων διαφοροποίησης ελάχιστα επηρεάζεται από τον τρόπο ορισμού των κρίσεων συνάφειας. Τα κριτήρια 'Συναίσθημα στις Ονοματικές Οντότητες' και 'Ονοματικές Οντότητες' παρουσιάζουν τις καλύτερες επιδόσεις (συμπεριλαμβανομένου του βασικού κριτήριο του περιεχομένου), (β)

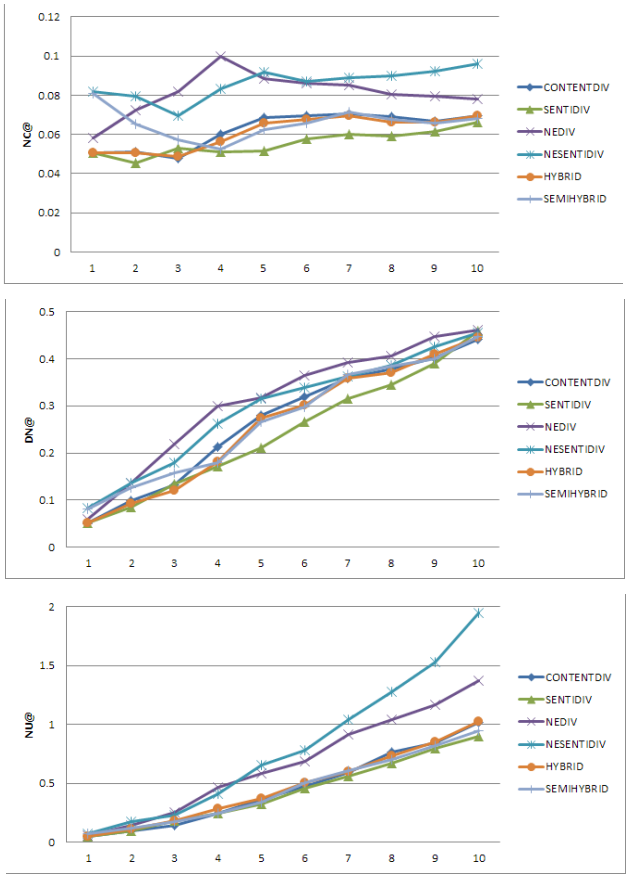


Σχήμα 6.10: Μέση αποτελεσματικότητα κάθε αλγορίθμου για όλες τις παραλλαγές. (Καλύτερη απεικόνιση έγχρωμα.)

Η σχετική αποτελεσματικότητα των αλγορίθμων διαφοροποίησης επηρεάζεται από την επιλογή των μονάδων πληροφοριών, ωστόσο, η συμπεριφορά του κάθε αλγόριθμου παραμένει η ίδια. Η συμπεριφορά αυτή πιθανός να οφείλεται στο ότι σημαντικές μονάδες πληροφορίας εντοπίζονται από τους χρήστες, αλλά όχι από τις αυτόματες μεθόδους και κατά συνέπεια μειώνεται η απόδοση τους, αλλά το σχήμα της καμπύλης απόδοσης παραμένει ίδιο και (γ) και στα δύο σενάρια αξιολόγησης, οι γραφικές παραστάσεις της μετρικής Distinct Nugget Coverage δείχνουν ότι υπάρχει ακόμα περιθώριο για βελτίωση τόσο των κριτηρίων όσο και των αλγορίθμων καθώς η μέγιστη κάλυψη των διακριτών μονάδων πληροφορίας που επιτυγχάνεται είναι 60%.

6.3 Διαφοροποιημένη Ανάκτηση Καταχωρήσεων σε ΜικροΙστολόγια

Σε αυτήν την ενότητα μελετάμε την διαφοροποιημένη ανάκτηση καταχωρήσεων σε κοινωνικά δίκτυα. Αρχικά καταδεικνύουμε την ανάγκη επέκτασης των τεχνικών διαφοροποίησης, ειδικά για το σενάριο της διαφοροποιημένης ανάκτησης σε κοινωνικά δίκτυα και οριοθετούμε το πρόβλημα. Στην συνέχεια εισάγουμε, με βάση τα ιδιαίτερα χαρακτηριστικά των καταχωρήσεων των χρηστών και των κοινωνικών δικτύων, εξειδικευμένα κριτήρια διαφοροποίησης



Σχήμα 6.11: Μέση αποτελεσματικότητα κάθε παραλλαγής για σε όλους τους αλγορίθμους. (Καλύτερη απεικόνιση έγχρωμα.)

και προσαρμόζουμε σε αυτά διαδεδομένους ευρετικούς αλγόριθμους διαφοροποίησης αποτελεσμάτων αναζήτησης. Τέλος, πραγματοποιούμε πειραματική αξιολόγηση των προαναφερθέντων μεθόδων και κριτηρίων διαφοροποίησης σε πραγματικά δεδομένα, μιας ιδιαίτερα δημοφιλούς υπηρεσίας μικροιστολογιών, του twitter.

6.3.1 Κίνητρο και Συνεισφορά

Τα κοινωνικά δίκτυα φιλοξενούν τεράστιες ποσότητες πληροφορίας και ο ρόλος τους για τη διάδοση πληροφοριών αυξάνεται σταθερά. Οι χρήστες των κοινωνικών δικτύων εκτός από το να δημοσιεύουν δικό τους περιεχόμενο αναζητούν πρόσφατες σχετικές πληροφορίες καθώς και πληροφορίες σχετιζόμενες με άλλους χρήστες. Η αναζήτηση στα κοινωνικά δίκτυα έχει αναδειχθεί ως μια νέα επιλογή για την κάλυψη των αναγκών των χρηστών, ιδίως όσον αφορά ειδήσεις ή τάσεις. Ταυτόχρονα, ο τεράστιος όγκος των καταχωρήσεων καθιστά αδύνατη την επισκόπηση του συνόλου των καταχωρήσεων από τους χρήστες, ενώ ο εντοπισμός χρήσιμης πληροφορίας σε ένα τεράστιο όγκο δεδομένων αποτελεί μια εξαιρετικά δύσκολη διεργασία. Συνεπώς αυτή η έκρηξη των δημοσιευμένων πληροφοριών θα μείνει ανεκμετάλλευτη εάν οι χρήστες δεν μπορούν να βρουν πληροφορίες που να παρέχουν ολοκληρωμένη κάλυψη των αναγκών τους.

Η διαφοροποίηση αποτελεσμάτων αναζήτησης, η οποία στοχεύει στη διευκόλυνση των χρηστών κατά την αναζήτηση πληροφορίας, συνίσταται στην αναταξινόμηση των αποτελεσμάτων και/ή στη συλλογή ενός περιορισμένου αριθμού αποτελεσμάτων, με τέτοιο τρόπο ώστε τα πρώτα αποτελέσματα που συλλέγονται να είναι όσο το δυνατόν πιο ετερογενή μεταξύ τους. Με τον τρόπο αυτό οι χρήστες θα ανακτήσουν αποτελέσματα τα οποία να καλύπτουν διαφορετικές οπτικές γωνίες της πληροφοριακής ανάγκης τους.

Η διαφοροποίηση βασισμένη μόνο στο κειμενικό περιεχόμενο των στοιχείων μπορεί να επαρκεί στο σενάριο της αναζήτησης εγγράφων με λέξεις κλειδιά, όπως αποδεικνύεται από τη βιβλιογραφία, δεν είναι όμως αρκετό, όταν πρόκειται για διαφοροποίηση καταχωρήσεων χρηστών σε κοινωνικά δίκτυα. Οι τεχνικές αναζήτησης και η ζητούμενη από τους χρήστες πληροφορία, διαφέρουν σε σχέση με το βασικό σενάριο αναζήτησης στο διαδύκτιο [138]. Για αυτό το λόγο, στο πλαίσιο που προτείνουμε, ορίζουμε εξειδικευμένα κριτήρια διαφοροποίησης που λαμβάνουν υπόψη τα χαρακτηριστικά τόσο καταχωρήσεων, όσο και του κοινωνικού δικτύου, με σκοπό να παράγουμε τις αντίστοιχες διαστάσεις διαφοροποίησης με τη μορφή διανυσμάτων χαρακτηριστικών. Στη συνέχεια, εφαρμόζουμε ευριστικούς αλγόριθμους διαφοροποίησης στις καταχωρήσεις των χρηστών, χρησιμοποιώντας τα παραπάνω κριτήρια. Το αποτέλεσμα της παραπάνω διαδικασίας είναι ένα υποσύνολο των αρχικών καταχωρήσεων που περιέχει ετερογενείς καταχωρήσεις, που αντιπροσωπεύουν διαφορετικές εκφάνσεις του προς εξέταση περιεχομένου, διαφορετικά είδη συναισθημάτων, διαφορετική ποιότητα γραφής, κτλ. Ταυτόχρονα, επεκτείνουμε υπάρχουσες μετρικές αξιολόγησης, προκειμένου να έχουν εφαρμογή στο υπό εξέταση σενάριο διαφοροποίησης. Πραγματοποιούμε πειραματική αξιολόγηση των μεθόδων μας, σε πραγματικά δεδομένα, δείχνοντας ότι τα κριτήρια που προτείνουμε επιφέρουν διακριτά πιο ετερογενή υποσύνολα διαφοροποιημένων καταχωρήσεων σε σύγκριση με τη βασική μέθοδο διαφοροποίησης μόνο με το κειμενικό περιεχόμενο.

Η συνεισφορά της εργασίας μας περιλαμβάνει τα ακόλουθα:

1. Ορίζουμε εξειδικευμένα κριτήρια διαφοροποίησης που λαμβάνουν υπόψη τα χαρακτηριστικά των καταχωρήσεων χρηστών σε κοινωνικά δίκτυα.

2. Εφαρμόσαμε, σε συνδυασμό με τις διαστάσεις διαφοροποίησης, τρεις ευριστικούς αλγόριθμους, ορίζοντας τις κατάλληλες συνθήκες αρχικοποίησης και συναρτήσεις ομοιότητας και συνάθροισης αποστάσεων.

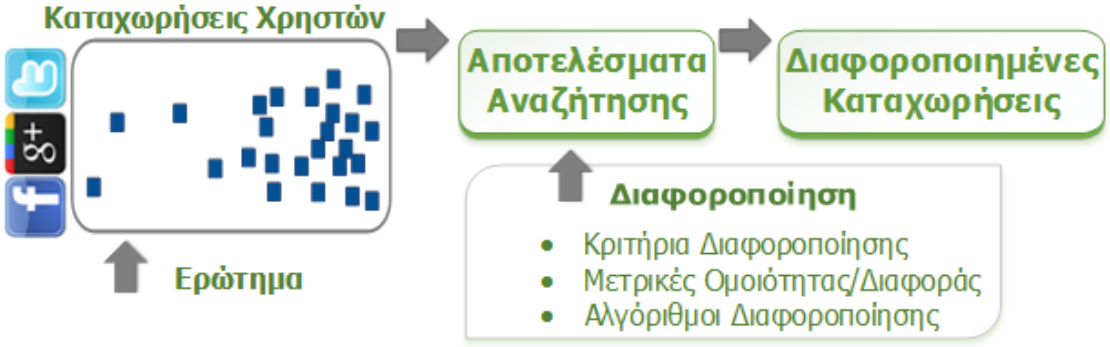
3. Αξιολογήσαμε πειραματικά την μεθόδων μας, χρησιμοποιώντας πραγματικά δεδομένα, δείχνοντας ότι τα κριτήρια που προτείνουμε επιφέρουν διακριτά πιο ετερογενή υποσύνολα διαφοροποιημένων καταχωρήσεων σε σύγκριση με τη βασική μέθοδο διαφοροποίησης μόνο με το κειμενικό περιεχόμενο.

6.3.2 Ορισμός Προβλήματος

Πιο συγκεκριμένα, το πρόβλημα ορίζεται ως εξής:

Ορισμός 6.1 (Διαφοροποίηση Νομικών Εγγράφων)**.** *Έστω q ένα ερώτημα του χρήστη και N ένα σύνολο από καταχωρήσεων χρηστών, σχετικές με το ερώτημα του χρήστη. Βρείτε*

ένα υποσύνολο $S \subset N$ καταχωρήσεων χρηστών που μεγιστοποιεί μία συνάρτηση στόχο f που ποσοτικοποιεί την ποικιλομορφία των θέσεων σε S.



Σχήμα 6.12: Διαδικασία Διαφοροποιημένης Ανάκτησης Καταχωρήσεων Χρηστών σε Κοινωνικά Δίκτυα

Μια επισκόπηση της μεθόδου που προτείνεται παρουσιάζεται στο Σχήμα 6.12. Για την εφαρμογή της μεθόδου θα πρέπει να οριστούν:

- **Κριτήρια Διαφοροποίησης**, ορίζουμε εξειδικευμένα σε καταχωρήσεις χρηστών σε κοινωνικά δίκτυα κριτήρια διαφοροποίησης με βάση τα οποία εκφράζουμε έκαστη καταχώρηση μέσω διανυσμάτων χαρακτηριστικών.
- **Συναρτήσεις αποστάσεων**, ορίζουμε συναρτήσεις μέτρησης της απόστασης μεταξύ δύο διανυσμάτων χαρακτηριστικών.
- **Αλγόριθμοι Διαφοροποίησης**, εφαρμόζουμε ευριστικούς αλγόριθμους διαφοροποίησης στις καταχωρήσεις χρηστών.

6.3.3 Κριτήρια Διαφοροποίησης

Οι καταχωρήσεις των χρηστών στα κοινωνικά δίκτυα (microblog posts), έχουν μοναδικά χαρακτηριστικά, που διαφοροποιούν την έρευνά μας από προηγούμενες εργασίες στη βιβλιογραφία. Πρώτον, το μέγιστο μήκος μιας καταχώρησης είναι μικρό. Προηγούμενες τεχνικές διαφοροποίησης χρησιμοποιούν προσεγγίσεις με έμφαση στην διαφοροποίηση μεγαλύτερου κειμένου, όπως ιστοσελίδες ή έγγραφα. Ομοίως, η γλώσσα που χρησιμοποιείται διαφέρει σημαντικά: δεδομένου ότι οι χρήστες αναρτούν μηνύματα για microblog υπηρεσίες από πολλά διαφορετικά μέσα, συμπεριλαμβανομένων των κινητών τηλεφώνων τους, η συχνότητα των ανορθογραφιών και συντμήσεων στις καταχωρήσεις τους είναι πάρα πολύ υψηλή σε σχέση με άλλους τομείς.

Με βάση τα προαναφερθέντα χαρακτηριστικά των καταχωρήσεων χρηστών σε κοινωνικά δίκτυα ορίζουμε τα ακόλουθα κριτήρια διαφοροποίησης:

- **Συναίσθημα**. Οι χρήστες, μέσω των μηνυμάτων τους, εκφράζουν τις απόψεις και το συναίσθημα τους για διάφορες οντότητες. Θεωρούμε ότι το συναίσθημα (θετικό,

αρνητικό ή ουδέτερο) είναι ένας παράγοντας διαφοροποίησης, δεδομένου ότι εκφράζει τις απόψεις των χρηστών. Υπό την έννοια αυτή, η ανάκτηση ενός συνόλου καταχωρήσεων που καλύπτουν διαφορετικές διαβαθμίσεις συναισθήματος και, κατά προτίμηση, με ομοιόμορφο τρόπο, ευνοεί την ποικιλομορφία/ετερογένεια.

- **Ονοματικές Οντότητες (named entities)**. Οι καταχωρήσεις των χρηστών στα κοινωνικά δίκτυα συχνά ακολουθούν την τρέχουσα επικαιρότητα και αφορούν συγκεκριμένες ονοματικές οντότητες (πρόσωπα, οργανισμούς και τοποθεσίες). Επομένως είναι σημαντικό ένα διαφοροποιημένο σύνολο καταχωρήσεων να περιλαμβάνει όσο το δυνατόν περισσότερες διαφορετικές ονοματικές οντότητες.

- **Γεωγραφική Θέση**. Μεταδεδομένα της αρχικής καταχώρησης περικλείουν την γεωγραφική θέση (γεωγραφικό μήκος και πλάτος) του χρήστη. Εφόσον η κοινωνική δραστηριότητα συσχετίζεται συχνά με γεωγραφικά όρια, η ανάκτηση ενός συνόλου καταχωρήσεων που καλύπτουν διαφορετικές γεωγραφικές περιοχές ευνοεί την ετερογένεια.

- **Χρόνος**. Το on-line περιεχόμενο αυξάνει και φθίνει με την πάροδο του χρόνου. Συνεπώς το χρονικό μοτίβο των καταχωρήσεων χρηστών παίζει σημαντικό ρόλο στην προσπάθεια για απόκτηση ευρέως σύνολου δημοσιεύσεων. Στόχος μας είναι να αποκτήσουμε καταχωρήσεις που να καλύπτουν ένα ευρύ χρονικό πλαίσιο, σε σχέση με το θέμα που σχετίζονται.

- **Κοινωνική επιρροή (Social Influence)**. Συμμετέχοντας σε κοινωνικά δίκτυα οι χρήστες μοιράζονται μια κοινή ανάγκη: να διαδώσουν τα μηνύματά τους σε όσο το δυνατόν περισσότερους χρήστες. Προτείνουμε ότι η κοινωνική επιρροή των χρηστών, η οποία ποσοτικοποιεί τη διάχυση των πληροφοριών σε μια υπηρεσία καταχωρήσεων, είναι ένας σημαντικός παράγοντας διαφοροποίησης των καταχωρίσεων.

- **Ποιότητα γραφής (Readability)**. Προτείνουμε ότι η ποιότητα γραφής είναι ένας παράγοντας διαφοροποίησης, δεδομένου ότι εκφράζει το βαθμό κατανόησης της καταχώρησης. Έτσι, είναι σημαντικό ένα σύνολο από καταχωρήσεις να περιλαμβάνει διαφορετικά επίπεδα καταληπτότητας - αναγνωσιμότητας. Η αναγνωσιμότητα μιας καταχώρησης υποδηλώνει το επίπεδο δυσκολίας του γραπτού κειμένου και προκύπτει από την εφαρμογή ενός τύπου αναγνωσιμότητας που λαμβάνει υπόψη ποιοτικά χαρακτηριστικά ενός κειμένου, όπως μήκος λέξεων και προτάσεων.

- **Κειμενικό περιεχόμενο**. Τέλος, εξετάζουμε το περιεχόμενο των καταχωρήσεων, το οποίο αποτελεί το βασικό (και πολλές φορές μόνο) κριτήριο διαφοροποίησης, που χρησιμοποιείται στις περισσότερες εργασίες που σχετίζονται με την διαφοροποίηση. Η σημασία του περιεχομένου των καταχωρήσεων στη διαδικασία διαφοροποίησης είναι προφανής.

6.3.4 Αλγόριθμοι Διαφοροποίησης

Οι αλγόριθμοι διαφοροποίησης που υλοποιήσαμε, Max-sum, Max-min, Mono-objective, έχουν παρουσιαστεί σε προηγούμενη εργασία [58] και περιγράφηκαν στην Ενότητα 5.2.4.

Έστω N ένα σύνολο από καταχωρήσεων χρηστών, σχετικές με το ερώτημα q του χρήστη, S το σύνολο διαφοροποιημένων καταχωρήσεων, $r(u,q)$ ο βαθμός ομοιότητας της καταχώρησης u με το ερώτημα q, $d(u,v)$ είναι η απόσταση μεταξύ των καταχωρήσεων u και v και $\lambda\epsilon[0,1]$ παράμετρος που προσδιορίζει το συμβιβασμό μεταξύ συνάφειας και ανομοιότητας.

– **Max-Sum**: στοχεύει στην μεγιστοποίηση του αθροίσματος της ομοιότητάς προς το ερώτημα και στην μεγιστοποίηση όλων των ανά δύο αποστάσεων μεταξύ όλων των στοιχείων του διαφοροποιημένου συνόλου S. Το σκορ για κάθε για κάθε υποψήφια καταχώρηση u, προκειμένου να επιλεγεί για εισαγωγή στο διαφοροποιημένο υποσύνολο S, μπορεί να εκφραστεί ως:

$$score_{MAXSUM}(u,v,q) = (1-\lambda)\cdot \frac{r(u,q)+r(v,q)}{2} + \lambda \cdot \sum_{i=1}^{|D|} w_i \cdot d_i(u,v) \qquad (6.13)$$

όπου i είναι το κριτήριο διαφοροποίησης, $|D|$ είναι ο αριθμός των κριτηρίων διαφοροποίησης και $w_i\epsilon[0,1]$ είναι το βάρος του κάθε επιμέρους diversity score, με $\sum_{i=1}^{|D|} w_i = 1$.

– **Max-Min**, στοχεύει στην μεγιστοποίηση της απόστασης μεταξύ των δύο πιο κοντινών σημείων μέσα στο διαφοροποιημένο, τελικό σύνολο καταχωρήσεων S. Μπορεί να εκφραστεί ως:

$$score_{MAXMIN}(u,q) = (1-\lambda)\cdot r(u,q) + \lambda \cdot \sum_{i=1}^{|D|} w_i \cdot d_i(u, \min v_{iu}) \qquad (6.14)$$

όπου $\min v_{iu}$ είναι η καταχώρηση από το τρέχον διαφοροποιημένο σύνολο με τη μικρότερη απόσταση από την υποψήφια καταχώρηση u

– **Mono-Objective**, στοχεύει στην ταυτόχρονη μεγιστοποίηση τόσο της ομοιότητας ενός στοιχείου με το ερώτημα, όσο και της απόστασης μεταξύ των καταχωρήσεων. Μπορεί να εκφραστεί ως:

$$score_{MONO}(u,q) = (1-\lambda)\cdot r(u,q) + \lambda \cdot \sum_{i=1}^{|D|} w_i \cdot \frac{1}{|N|-1} \sum_{v\epsilon N} d(u,v) \qquad (6.15)$$

6.3.5 Συναρτήσεις Αποστάσεων

Στην Ενότητα 6.3.3 περιγράφηκαν τα κριτήρια διαφοροποίησης, με βάση τα οποία εκφράζουμε έκαστη καταχώρηση μέσω διανυσμάτων χαρακτηριστικών, που αντιστοιχούν στα κριτήρια διαφοροποίησης και τα οποία αντιπροσωπεύουν διάφορες εκφάνσεις των καταχωρήσεων. Οι αλγόριθμοι διαφοροποίησης, Ενότητα 6.3.4, χρησιμοποιούν αυτά τα διανύσματα για

να υπολογίζουν, σε κάθε βήμα, ένα αθροιστικό σκορ διαφοροποίησης για κάθε καταχώρηση. Αυτό το σκορ, στη συνέχεια, σταθμίζεται με το σκορ ομοιότητας της καταχώρισης με το αρχικό ερώτημα του χρήστη για να παραγάγει ένα τελικό σκορ, από το οποίο καθορίζεται η επιλογή της επόμενης καταχώρισης. Προκειμένου να παράγουμε σκορ διαφοροποίησης, πρέπει να ορίσουμε μία συνάρτηση η οποία θα μετράει την απόσταση μεταξύ δύο στοιχείων.

Για τα κριτήρια διαφοροποίησης που εκφράζονται σε μορφή διανυσμάτων χαρακτηριστικών, *[Συναίσθημα, Ονομαστικές Οντότητες, Κειμενικό περιεχόμενο]*, χρησιμοποιούμε την συνάρτηση ομοιότητας συνημίτονου (cosine similarity function), σε κανονικοποιημένη μορφή, και ορίζουμε το σκορ διαφοροποίησης μεταξύ δύο στοιχείων, u, v, με διανυσμάτων χαρακτηριστικών, $V(u), V(v)$, ως προς μία διάσταση-κριτήριο διαφοροποίησης i, ως εξής:

$$d_i(u, v) = 1 - \cos_i(V(u), V(v)) \tag{6.16}$$

Για τα κριτήρια διαφοροποίησης που εκφράζονται σε μορφή αριθμητικών τιμών, *[Γεωγραφική Θέση, Χρόνος, Οντότητες, Κοινωνική επιρροή, Ποιότητα γραφής]*, ορίζουμε το σκορ διαφοροποίησης μεταξύ δύο στοιχείων, u, v, με κανονικοποιημένες στο [0..1] τιμές χαρακτηριστικών, $N(u), N(v)$, ως προς μία διάσταση-κριτήριο διαφοροποίησης i, ως εξής:

$$d_i(u, v) = 1 - |N_i(u) - N_i(v)| \tag{6.17}$$

Σημειώνουμε ότι η κανονικοποίηση γίνεται στο επίπεδο του κάθε κριτηρίου ξεχωριστά. Δηλαδή, υπολογίζουμε τη μέγιστη τιμή που μπορεί να πάρει κάποιο κριτηρίου για όλες τις καταχωρήσεις και διαιρούμε τα αντίστοιχα σκορ των υπολοίπων καταχωρήσεων με αυτό, για κάθε κριτήριο ξεχωριστά.

Η ετερογένεια μεταξύ των καταχωρήσεων δεν είναι ο μόνος στόχος: οι καταχωρήσεις θα πρέπει να είναι, επιπλέον, σχετικές με το ερώτημα του χρήστη. Έτσι, το τελικό σκορ για κάθε καταχώρηση είναι το σταθμισμένο άθροισμα του συνολικού σκορ διαφοροποίησης του και του σκορ ομοιότητάς του με το ερώτημα. Το σκορ ομοιότητας r μιας καταχώρησης u με ένα ερώτημα q, ορίζεται με βάση τη συνάρτηση συνημίτονου στα διανύσματα χαρακτηριστικών όρων της καταχώρησης και του ερωτήματος:

$$r(u, q) = \cos(V(u), V(q)) \tag{6.18}$$

Σημειώνουμε ότι και αυτό το σκορ κανονικοποιείται στο διάστημα [0..1].

6.4 Πειραματική Μελέτη

Στην ενότητα αυτή, περιγράφουμε τη συλλογή καταχωρήσεων χρηστών που χρησιμοποιήσαμε, τις μετρικές που χρησιμοποιούμε για την αξιολόγηση, τις κατηγορίες ερωτημάτων, την μεθοδολογία που σχεδιάσαμε για την αντικειμενική επισημείωση με κρίσεις συνάφειας των καταχωρήσεων για κάθε ερώτημα, καθώς τα σενάρια διαφοροποίησης που αξιολογούμε. Τέλος, παρέχουμε τα αποτελέσματα μαζί με μια σύντομη συζήτηση.

6.4.1 Σενάρια Αξιολόγησης

Αξιολογήσαμε τα ακόλουθα σενάρια, χρησιμοποιώντας συνδυασμό κριτηρίων διαφοροποίησης ως:

- **Κειμενική διαφοροποιηση - CONTENTDIV**: Η βασική μέθοδος σύγκρισης που εφαρμόζει διαφοροποίηση μόνο πάνω στο κειμενικό περιεχόμενο.
- **Διαφοροποίηση ονοματικών οντοτήτων - NEDIV**: Η παραλλαγή που διαφοροποιεί μόνο πάνω στις αναγνωριζόμενες ονοματικές οντότητες.
- **Διαφοροποίηση συναισθήματος - SENTIDIV**: Η παραλλαγή που διαφοροποιεί μόνο πάνω στο αναγνωριζόμενο συναίσθημα των σχολίων.
- **Διαφοροποίηση συναισθήματος ονοματικών οντοτήτων - NESENTIDIV**: Η παραλλαγή που διαφοροποιεί πάνω στο συναίσθημα που αναγνωρίζεται γύρω από τις αναγνωριζόμενες ονοματικές οντότητες του σχολίου.
- **Χρονική διαφοροποίηση - TIMEDIV**: Η παραλλαγή που διαφοροποιεί χρησιμοποιώντας την χρονική συνιστώσα των καταχωρήσεων.
- **Χωρική διαφοροποίηση - PLACEDIV**: Η παραλλαγή που διαφοροποιεί χρησιμοποιώντας την γεωγραφική συνιστώσα των καταχωρήσεων.
- **Διαφοροποίηση Ποιότητας γραφής - READABILITYDIV**: Η παραλλαγή που διαφοροποιεί χρησιμοποιώντας την ποιότητα γραφής των καταχωρήσεων.
- **Διαφοροποίηση Κοινωνικής επιρροής - READABILITYDIV**: Η παραλλαγή που χρησιμοποιεί την κοινωνική επιρροή των χρηστών ως παράγοντα διαφοροποίησης των καταχωρίσεων.
- **Υβριδική διαφοροποίηση - SEMIHYBRID**: Η παραλλαγή που διαφοροποιεί συνδυάζοντας τα κριτήρια της κειμενικής ομοιότητας, του συναισθήματος, των ονοματικών οντοτήτων, του χρόνου και γεωγραφικής Θέσης, με ισοκατανεμημένα τα επί μέρους βάρη των κριτηρίων.
- **Εκτεταμένη υβριδική διαφοροποίηση - HYBRID**: Η παραλλαγή που διαφοροποιεί συνδυάζοντας όλα προτεινόμενα κριτήρια, με ισοκατανεμημένα τα επί μέρους βάρη των κριτηρίων.

Κάθε ερώτημα χρήστη υποβάλλεται στο σύστημα και ανακτώνται τα $top - 100$ έγγραφα, τα οποία σχηματίζουν το υποψήφιο σύνολο N, με χρήση της εξίσωσης 5.5 και το log-based $tf - idf$ σχήμα ευρετηρίασης. Η επιλεγμένη τιμή, $n = 100$, για το πλήθος του υποψηφίου συνόλου N είναι μια τυπική τιμή που χρησιμοποιείται στη βιβλιογραφία [126]. Στην συνέχεια κάθε ένα από τα σενάρια διαφοροποίησης εφαρμόζεται σε συνδυασμό με κάθε ένα από τους αλγορίθμους διαφοροποίησης και υποψήφιο σύνολο N, άρα για κάθε ερώτημα του χρήστη. Για

όλες τις μεθόδους διαφοροποίησης η παράμετρος trade-off λ ρυθμίζεται σε σταθερή τιμή 0.5 και επομένως $\lambda = 0.5 = 1 - \lambda$, δηλαδή επιδιώκουμε ισοκατανομή των βαρών της συνάφειας και της διαφορετικότητας στο τελικό αποτέλεσμα. Σημειώνουμε τέλος ότι στις μεθόδους που συνδυάζουν κριτήρια, τα σκορ των επιμέρους κριτηρίων σταθμίζονται ισοδύναμα ώστε να παραχθεί το τελικό σκορ διαφοροποίησης.

6.4.2 Συλλογή Καταχωρήσεων Χρηστών

Για την αξιολόγηση χρησιμοποιήσαμε ένα σύνολο καταχωρήσεων χρηστών. Αυτές ανακτήθηκαν από το twitter χρησιμοποιώντας την ροή δεδομένων του, twitter API, κατά τη διάρκεια του διαγωνισμού τραγουδιού της Eurovision και περιέχουν την λέξη κλειδί 'eurovision'. Οι καταχωρήσεις χρηστών στα κοινωνικά δίκτυα, όπως προαναφέρθηκε, έχουν ιδιαίτερη σύνταξη γεγονός που δυσχεραίνει την διαδικασία ανάκτησης από τα υπάρχοντα μοντέλα ανάκτησης πληροφοριών. Έτσι είναι απαραίτητη μια προ-επεξεργασία των καταχωρήσεων, η οποία συνίσταται στην αφαίρεση συγκεκριμένων συμβόλων, πχ #, @, url και των μη λατινικών χαρακτήρων. Επιπρόσθετα, για την εφαρμογή του κριτηρίου διαφοροποίησης 'Γεωγραφική Θέση' αφαιρέσαμε τις μη γεωγραφικά επισημειωμένες καταχωρήσεις. Το ευρετήριο μας κατασκευάστηκε χρησιμοποιώντας τυπική λίστα κοινών λέξεων και τεχνική Porter stemming, με log-based $tf-idf$ σχήμα ευρετηρίασης, έχοντας συνολικά 24,000 καταχωρήσεις χρηστών. Για την εξαγωγή των ονοματικών οντοτήτων χρησιμοποιήσαμε το Stanford Named Entity[7] [44], του συναίσθηματος το Sentistrength [139] και της κοινωνικής επιροής το InfluenceTracker[8] [118]. Τέλος, υπολογίστηκαν και κανονικοποιήθηκαν οι τιμές των υπόλοιπων κριτηρίων διαφοροποίησης, Γεωγραφική Θέση, Χρόνος και Ποιότητα γραφής, για κάθε καταχώρηση χρήστη της συλλογής δεδομένων μας.

Σημειώνουμε επίσης ότι, η λέξη-κλειδί που χρησιμοποιήσαμε είναι ταυτόχρονα αρκετά γενική, ώστε να μας εξασφαλίσει ικανοποιητικό αριθμό καταχωρήσεων και αρκετά 'ειδική', ώστε οι καταχωρήσεις αυτές είναι πλούσιες τόσο σε ονοματικές οντότητες, όσο και στην έκφανση του συναισθήματος των χρηστών σε αυτές.

6.4.3 Μετρικές Αποτίμησης

Οι μετρικές αποτίμησης και η γενικότερη μεθοδολογία αξιολόγησης που χρησιμοποιούμε βασίζονται στην έννοια των μονάδων πληροφορίας, δηλαδή κάθε έννοια ή θεματική κατηγορία ή υποκατηγορία ή σχετική άποψη / συναίσθημα ή επέκταση των παραπάνω εννοιών και κατηγοριών που σχετίζεται με την καταχώρηση του χρήστη.

Οι μετρικές αποτίμησης που χρησιμοποιούμε είναι:

– **Nugget Coverage - NC@n**: εκφράζει την κάλυψη των μονάδων πληροφορίας στο σύνολο των αποτελεσμάτων. (Εξίσωση 6.8)

[7] `http://nlp.stanford.edu/software/CRF-NER.shtml`
[8] `http://influencetracker.com/`

– **Distinct Nugget Coverage - DN@n**: ο λόγος των διακριτών μονάδων πληροφορίας που βρίσκονται στις καταχωρήσεις, προς το συνολικό αριθμό διακριτών μονάδων. (Εξίσωση 6.10)

– **Nugget Uniformity - NU@n**: ποσοτικοποιεί την διακύμανση των μονάδων πληροφορίας στο σύνολο αποτελεσμάτων. (Εξίσωση 6.12)

6.4.4 Κρίσεις Συνάφειας

Όπως αναφέρθηκε και στην Ενότητα 5.3.3, μια από τις δυσκολίες στην αξιολόγηση μεθόδων που έχουν σχεδιαστεί για την διαφοροποίηση αποτελεσμάτων αποτελεί η έλλειψη τυπικών δεδομένων δοκιμών , η έλλειψη ενός συνόλου αντικειμενικής αλήθειας. Στην περίπτωση μας, διαθέτοντας μόνο την συλλογή καταχωρήσεων χρηστών, χρειάστηκε να καθορίσουμε επιπρόσθετα: (α) τις κατηγορίες ερωτημάτων, (β) μια μέθοδο για τον προσδιορισμό των επιμέρους θεμάτων σε κάθε θέμα και (γ) μια μέθοδο για την υποσημείωση των καταχωρήσεων χρηστών με κρίσεις συνάφειας για το κάθε θέμα.

Με βάση την έλλειψη τυπικών δοκιμαστικών δεδομένων, ερευνήσαμε για ένα αντικειμενικό τρόπο για εκτιμήσουμε και να αξιολογήσουμε τις επιδόσεις των διαφόρων μεθόδων διαφοροποίησης στην συλλογή καταχωρήσεων μας. Αναγνωρίζουμε το γεγονός ότι η διαδικασία της αυτόματης παραγωγής ερωτημάτων και κρίσεων συνάφειας αποτελεί, στην καλύτερη των περιπτώσεων, μια ατελή προσέγγιση των ενεργειών ενός πραγματικού χρήστη, όμως ο τεράστιος αριθμός των καταχωρήσεων στα κοινωνικά δίκτυα καθιστά ιδιαίτερα δαπανηρή την διαδικασία ορισμού από τον χρήστη των κρίσεων συνάφεια.

Προφίλ χρηστών / ερωτήματα

Εκπαιδεύσαμε, στην συλλογή δεδομένων μας, ένα μοντέλο με βάση τον αλγόριθμο λανθάνουσας κατανομής Dirichlet - LDA [19] (LDA topic model), χρησιμοποιώντας την εφαρμογή ανοικτού λογισμικού Mallet[9]. Η μοντελοποίηση αυτή μας παρέχει έναν τρόπο να συναχθεί η λανθάνουσα δομή πίσω από μια συλλογή εγγράφων αφού κάθε παραγόμενο από την μέθοδο θέμα αποτελείται από ένα σύνολο από λέξεις κλειδιά με αντίστοιχα βάρη. Απαιτήσαμε από την μέθοδο την παραγωγή 10 θεμάτων, εκ των οποίων κρατήσαμε τις 20 κορυφαίες λέξεις-κλειδιά για κάθε θέμα. Στην συνέχεια 2 ερευνητές του εργαστηρίου μας, δημιούργησαν ένα ερώτημα χρήστη χρησιμοποιώντας ένα τυχαίο αριθμό από τις λέξεις κλειδιά για κάθε θέμα. Τα επιλεγμένα ερωτήματα για να αντιπροσωπεύουν τις ανάγκες πληροφόρησης των χρηστών παρουσιάζονται στον Πίνακα 6.9.

[9] `http://mallet.cs.umass.edu/`

Πίνακας 6.9: Επιλεχθέντα ερωτήματα χρηστών

russia europe	song contest
conchita	eurovision final
watch	poland girls
party time	sing
eurovison winner	tonight

Επισημειώσεις ερωτημάτων και Σύνολο Αντικειμενικής Αλήθειας

Υποβάλαμε κάθε ένα από τα ερωτήματα χρηστών, στο σύστημα ανάκτησης μας, το οποίο μας επέστρεψε τις σχετικές καταχωρήσεις χρηστών με βάση την συνάφεια με το ερώτημα, χρησιμοποιώντας την εξίσωση ομοιότητας καταχώρησης με ερώτημα χρήστη (Εξίσωση 6.18). Εν συνεχεία για κάθε ένα από τα $top - n$ αποτελέσματα, για κάθε ερώτηση, χρησιμοποιήσαμε την υπηρεσία OpenCalais Web Service[10], μια υπηρεσία η οποία επιστρέφει σημασιολογικά μεταδεδομένα για κάθε τεθέν κείμενο. Τα μεταδεδομένα αυτά εν συνεχεία τα θεωρούμε ως μονάδες πληροφορίας τόσο για κάθε προκύπτουσα καταχώρηση, όσο και συγκεντρωτικά, για το έκαστο ερώτημα χρήστη. Με τον τρόπο αυτό, στηριζόμενοι σε αντικειμενικές μεθόδους, επισημειώσαμε την συλλογή καταχωρήσεων μας με κρίσεις μονάδων πληροφορίας.

Ο Πίνακας 6.10 παρέχει ένα δείγμα από τυχαίες καταχωρήσεις χρηστών και προσπίπτουσες μονάδες πληροφορίας για το Ερώτημα : party time.

Πίνακας 6.10: Δείγμα από τυχαίες καταχωρήσεις χρηστών και προσπίπτουσες μονάδες πληροφορίας για το Ερώτημα : party time

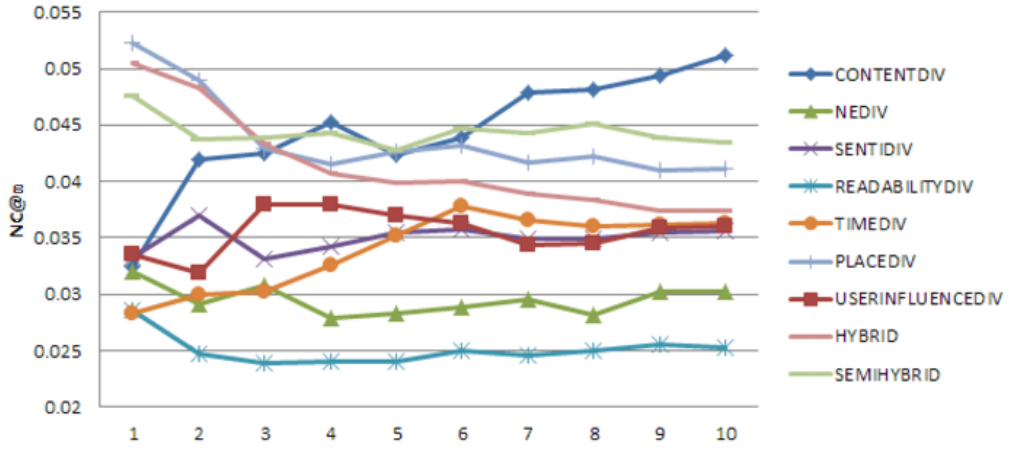
Tweet	**Indicative nuggets**
@my_space oh no ha ha. I have fallen into the trap of palms. Throw a Eurovision party on Sunday for us	human interest, politics, eurovision party
Eurovision party is over boo (at @CrownMetropol Hotel) http://t.co/E0m5SB5NO2	CrownMetropol hotel, eurovision
Feeding time #JoinUs #Eurovision #Copengaygen #Copenhagen @Royal Bagel Copenhagen http://t.co/Ko7ihngPvI	human interest, copenhagen, food and drink, eurovision, bagel
5Now, #Eurovision party over, time to head home! (@Gants Hill London Underground Station - @tflofficial) http://t.co/XNpwbqXheJ	hospitality_recreation, law_crime, eurovision, london borough of redbridge, london, gants hill, the gants, ilford

6.4.5 Αποτελέσματα

Ως βάση για την σύγκριση των μεθόδων διαφοροποίησης, θεωρούμε την καθιερωμένη στο σενάριο της διαφοροποίησης αποτελεσμάτων αναζήτησης μέθοδο Content Diversity - CONTENTDIV, η οποία διαφοροποιεί τις καταχωρήσεις χρηστών βασιζόμενη μόνο στο κριτήριο της κειμενικής ομοιότητας. Για κάθε ερώτημα, το αρχικό σύνολο N περιέχει τα $top - 100$ αποτελέσματα. Τα αποτελέσματα παρουσιάζονται με σταθερή παράμετρο $n = |N|$. Σημειώνουμε ότι κάθε μία από τις δοκιμές διαφοροποίησης εφαρμόζεται σε συνδυασμό με κάθε ένα

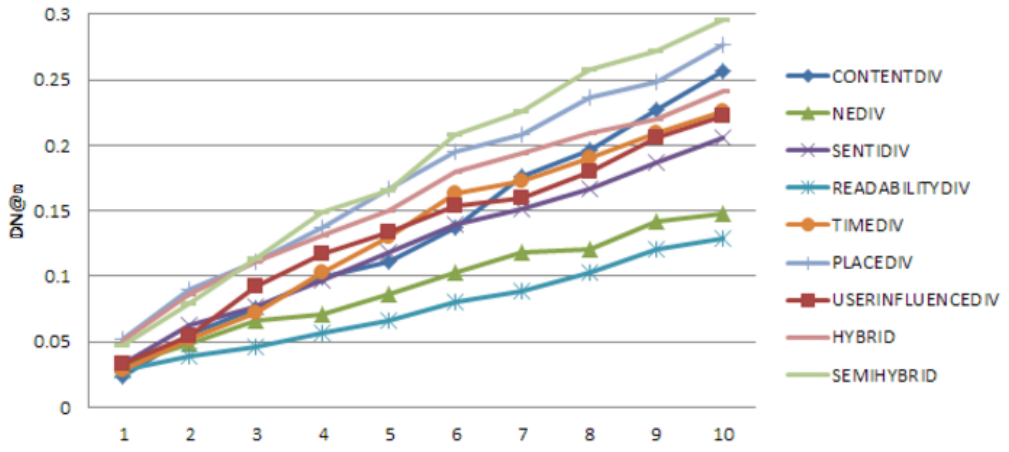
[10] `http://www.opencalais.com/`

από τους αλγορίθμους διαφοροποίησης και για κάθε ερώτημα του χρήστη.



Σχήμα 6.13: Μετρήσεις Nugget Coverage - NC@n στις θέσεις 1 έως 10, για μέτρηση βάσης CONTENTDIV και σενάρια NEDIV, SENTIDIV, READABILITYDIV, TIMEDIV, PLACEDIV, USERINFLUENCEDIV, HYBRIDDIV και SEMIHYBRID.(Καλύτερη απεικόνιση έγχρωμα.)

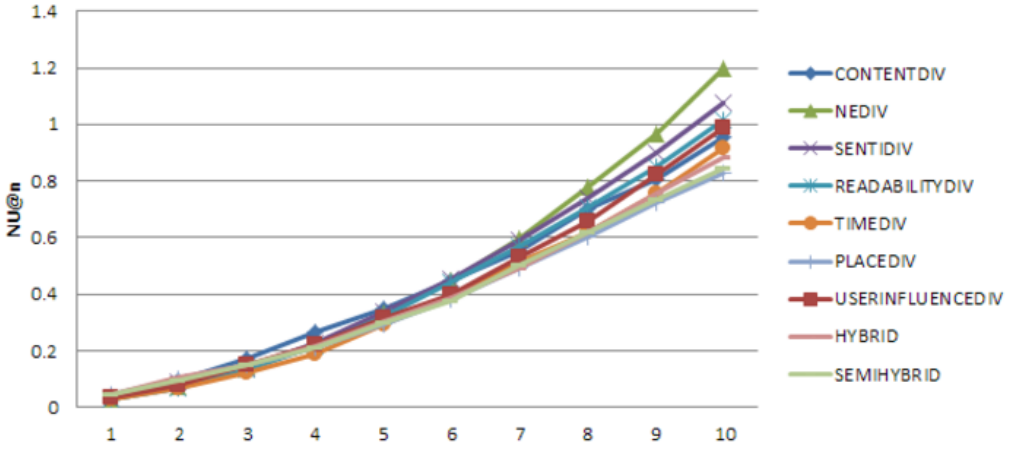
Στο Σχήμα 6.13, παρουσιάζεται η μετρική Nugget Coverage - NC@n στις θέσεις 1 έως 10. Σημειώνουμε ότι οι μετρικές Nugget Coverage και Distinct Nugget Coverage κανονικοποιούνται, εξ ορισμού, σε το διάστημα [0..1]. Παρατηρούμε ότι η παραλλαγή SEMIHYBRID επιφέρει καλύτερα αποτελέσματα μέχρι την θέση 6 και οι παραλλαγές Εκτεταμένη υβριδική διαφοροποίηση - HYBRID και Χωρική διαφοροποίηση - PLACEDIV μέχρι την θέση 3, σε σχέση με την βάση αξιολόγησης, την Κειμενική διαφοροποίηση - CONTENTDIV. Η απόδοση των παραλλαγών αυτών εμφανίζει σταδιακή πτώση, σε αντίθεση με την παραλλαγή CONTENTDIV, οι τιμές της οποίας είναι γενικά αυξανόμενες.



Σχήμα 6.14: Μετρήσεις Distinct Nugget Coverage - DN@n στις θέσεις 1 έως 10, για μέτρηση βάσης CONTENTDIV και σενάρια NEDIV, SENTIDIV, READABILITYDIV, TIMEDIV, PLACEDIV, USERINFLUENCEDIV, HYBRIDDIV και SEMIHYBRID.(Καλύτερη απεικόνιση έγχρωμα.)

Στο Σχήμα 6.14, απεικονίζει τις τιμές της μετρικής Distinct Nugget Coverage στις θέσεις

1 έως 10. Σε αντίθεση με την μετρική Nugget Coverage, παρατηρούμε ότι οι παραλλαγές Υβριδικής διαφοροποίησης SEMIHYBRID, η οποία συνδυάζει τα κριτήρια της κειμενικής ομοιότητας, του συναισθήματος, των ονοματικών οντοτήτων, του χρόνου και γεωγραφικής Θέσης, και Χωρικής διαφοροποίησης - PLACEDIV επιφέρουν διακριτά καλύτερα αποτελέσματα σε σχέση με την μέτρηση βάσης CONTENTDIV και την παραλλαγή που συνδυάζει όλα τα κριτήρια (HYBRID) που ακολουθούν. Τέλος, οι υπόλοιπες υπό εξέταση παραλλαγές έπονται. Δεδομένου ότι οι μετρήσεις Distinct Nugget Coverage αφορούν διακριτές μονάδες πληροφορίας των πληροφοριών που εμπεριέχονται στις καταχωρήσεις των χρηστών, τα αποτελέσματα αυτά, παρέχουν μια ισχυρή ένδειξη της διαφορετικότητας. Ταυτόχρονα οι παραλλαγές διαφοροποίηση ποιότητας γραφής - READABILITYDIVκαι διαφοροποίηση ονοματικών οντοτήτων - NEDIV αποτυγχάνουν να βελτιώσουν την μέτρηση βάσης, επιφέροντας χειρότερα αποτελέσματα από την αρχική τιμή κατάταξης σε όλα τα επίπεδα. Αποδίδουμε το εύρημα αυτό στο πολύ μικρό μήκος των καταχωρήσεων χρηστών σε μικροιστολόγια, μέγιστο 140 χαρακτήρες, που δεν επιτρέπει την αποτελεσματική λειτουργιά των αντίστοιχων συναρτήσεων.



Σχήμα 6.15: Μετρήσεις DNugget Uniformity - NU@n στις θέσεις 1 έως 10, για μέτρηση βάσης CONTENTDIV και σενάρια NEDIV, SENTIDIV, READABILITYDIV, TIMEDIV, PLACEDIV, USERINFLUENCEDIV, HYBRIDDIV και SEMIHYBRID.(Καλύτερη απεικόνιση έγχρωμα.)

Στο Σχήμα 6.15, απεικονίζει τις τιμές της μετρικής Nugget Uniformity στις θέσεις 1 έως 10. Η μετρική Nugget Uniformity ποσοτικοποιεί την διακύμανση των μονάδων πληροφορίας στο σύνολο αποτελεσμάτων και σε αντίθεση με τις προαναφερθείσες μετρικές χαμηλότερες τιμές σημαίνουν καλύτερη απόδοση. Ομοίως οι παραλλαγές Υβριδικής διαφοροποίησης SEMIHYBRID και Χωρικής διαφοροποίησης - PLACEDIV επιφέρουν διακριτά καλύτερα αποτελέσματα σε σχέση με την μέτρηση βάσης CONTENTDIV και τις υπόλοιπες παραλλαγές. Συνολικά η ανομοιογένεια των μονάδων πληροφορίας, διανέμεται καλύτερα μέσω του συνδυασμού των κριτηρίων της κειμενικής ομοιότητας, του συναισθήματος, των ονοματικών οντοτήτων, του χρόνου και γεωγραφικής Θέσης.

Γενικά, παρατηρούμε ότι η χρήση εξειδικευμένων κριτηρίων διαφοροποίησης, σε σχέση με την απλή κειμενική ομοιότητα, επιφέρει σημαντικές βελτιώσεις στην κάλυψη των μονάδων

πληροφορίας που αντιπροσωπεύουν την πληροφοριακή ανάγκη ενός χρήστη και κατά συνέπεια στην αποτελεσματικότητα της διαδικασίας διαφοροποίησης στις καταχωρήσεις χρηστών στα κοινωνικά δίκτυα.

6.5 Συμπεράσματα

Στο κεφάλαιο αυτό μελετήθηκε το πρόβλημα της διαφοροποιημένης ανάκτησης καταχωρήσεων σε κείμενα διαβουλεύσεων και σε κοινωνικά δίκτυα.

Τεχνικές διαφοροποίησης αποτελεσμάτων αναζήτησης, όπως η διαφοροποίηση βασιζόμενη μόνο στο κειμενικό περιεχόμενο ενός πόρου, οι οποίες λειτουργούν αποτελεσματικά με βάση τα χαρακτηριστικά του παγκοσμίου ιστού, είναι ανεπαρκείς στο σενάριο που εξετάζουμε. Ορίσαμε εξειδικευμένα κριτήρια διαφοροποίησης, τα οποία αποτυπώνουν αποτελεσματικότερα τις διάφορες πτυχές/ εκφάνσεις των πληροφοριακών αναγκών των χρηστών. Στα κριτήρια αυτά προσαρμόσαμε διαδεδομένους ευρετικούς αλγορίθμους, καθώς και μια παραλλαγή αλγόριθμου διαφοροποίησης που αποδίδει πολύ κοντά στον βέλτιστο δοκιμαζόμενο αλγόριθμο. Για τις ανάγκες αξιολόγησης της αποτελεσματικότητας των μεθόδων, ορίσαμε μετρικές, επεκτείνοντας την έννοια των μονάδων πληροφορίας. Η πειραματική αξιολόγηση των μεθόδων, που πραγματοποιήθηκε με βάση δημοσίως διαθέσιμα πραγματικά σύνολα δεδομένων, κατέδειξε την υπεροχή της μεθόδου μας, σε σχέση με τις υφιστάμενες προσεγγίσεις και την διαφοροποίηση μόνο του κειμενικού περιεχομένου των πόρων που χρησιμοποιούν.

Σε μελλοντικές εργασίες, σχεδιάζουμε επίσης να μελετήσουμε περαιτέρω την αλληλεπίδραση του κριτηρίου 'Συν-σχολιασμός Χρηστών', εφαρμόζοντας τεχνικές θεματικής συσταδοποίησης για να αποκτήσουμε ένα πιο πυκνό διάνυσμα χαρακτηριστικών. Ταυτόχρονα επιθυμούμε να αξιολογήσουμε την απόδοση υβριδικών συνδυασμών αλγορίθμων.

Κεφάλαιο 7

Σύνοψη και Μελλοντικές Επεκτάσεις

Στο τελευταίο κεφάλαιο της διατριβής συνοψίζουμε τη συνεισφορά μας και συζητάμε προτάσεις για μελλοντικές εργασίες πάνω στα θέματα που πραγματευτήκαμε στο παρόν.

7.1 Σύνοψη

Στα πλαίσια της διατριβής εστιάσαμε το ενδιαφέρον μας σε προβλήματα και προκλήσεις του κλάδου της νομικής πληροφορικής. Αρχικά πραγματοποιήσαμε μια εκτενή βιβλιογραφική μελέτη για όλες τις σχετικές ερευνητικές εργασίες.

Στην συνέχεια προτείναμε την αρχιτεκτονική μιας πλατφόρμας διαχείρισης νομικών πηγών. Στόχος της είναι η βελτίωση της πρόσβασης σε νομικές πηγές παρέχοντας προηγμένες υπηρεσίες μοντελοποίησης, διαχείρισης και ανάκτησης νομικής πληροφορίας. Χρησιμοποιεί μια πρότυπη μέθοδο για την εξαγωγή σημασιολογικών αναπαραστάσεων νομικών πηγών από μη δομημένες μορφές, μέσω της δημιουργίας μιας γλώσσας συγκεκριμένου τομέα για τις νομικές πηγές. Η αρχιτεκτονική αξιολογήθηκε σε περιβάλλον παραγωγής, δημόσιου φορέα, παρέχοντας στο ευρύ κοινό σημασιολογική πρόσβαση στην ελληνική φορολογική νομοθεσία.

Παράλληλα, προτείναμε μια καινοτόμο προσέγγιση για την μοντελοποίηση του δικαίου, σε μορφή σύνθετου δικτύου, ενός δικτύου πολλαπλών σχέσεων που φιλοξενεί την ιεραρχία μεταξύ των πηγών του δικαίου και μπορεί να αντιπροσωπεύει σχέσεις διαφόρων κατηγοριών μεταξύ νομικών πηγών, μαζί με τη χρονική εξέλιξή τους. Εφαρμόσαμε το μοντέλο στο δίκαιο της Ε.Ε αναλύοντας την τοπολογία του, μελετώντας τη χρονική του εξέλιξη και αξιολογώντας την ανθεκτικότητα του σε μεταβολές. Η ανάλυση μας κατέδειξε αφανείς οργανωτικές αρχές του σώματος του δικαίου.

Επίσης, εξετάσαμε τη μεγιστοποίηση της νομικής ποικιλομορφίας των αποτελεσμάτων αναζήτησης. Προς την κατεύθυνση αυτή, προσαρμόσαμε αλγορίθμους που έχουν προταθεί για την κάλυψη ετερογενών αναγκών π.χ., την δημιουργία περιλήψεων, την κατάταξη σε γράφους και την διαφοροποίηση αποτελεσμάτων αναζήτησης. Προτείναμε επίσης εξειδικευμένα κριτήρια διαφοροποίησης νομικών πηγών τα οποία και ενσωματώσαμε στους αλγορίθμους. Αξιολο-

γήσαμε σε πραγματικές συλλογές νομικών εγγράφων την απόδοση των αλγορίθμων και των κριτηρίων, προσφέροντας όρια εξισορρόπησης για τα συστήματα ανάκτησης νομικής πληροφορίας, που επιθυμούν να ισορροπήσουν μεταξύ της ενίσχυσης των σχετικών εγγράφων ή να δειγματοληπτήσουν τον χώρο νομικής πληροφορίας γύρω από το ερώτημα.

Τέλος, μελετήσαμε τη διαφοροποιημένη ανάκτηση καταχωρήσεων σε κείμενα διαβουλεύσεων και σε κοινωνικά δίκτυα. Ορίσαμε κριτήρια διαφοροποίησης που λαμβάνουν υπόψη τα χαρακτηριστικά των καταχωρήσεων και του κοινωνικού δικτύου. Ενσωματώσαμε τα κριτήρια αυτά σε ευριστικούς αλγόριθμους διαφοροποίησης και αξιολογήσαμε, με βάση δημοσίως διαθέσιμα πραγματικά σύνολα δεδομένων, την απόδοση των κριτηρίων και αλγορίθμων διαφοροποίησης. Τα αποτελέσματα κατέδειξαν την υπεροχή των κριτηρίων διαφοροποίησης, σε σχέση με τις υφιστάμενες προσεγγίσεις.

7.2 Μελλοντικές εργασίες

Κατά την εκπόνηση της παρούσας διατριβής, αναγνωρίσαμε τα ακόλουθα ενδιαφέροντα θέματα τα οποία προτείνουμε για μελλοντική εργασία.

Στην κατεύθυνση της μοντελοποίησης και διαχείρισης των νομικών πηγών διάφορες επεκτάσεις βρίσκονται υπό διερεύνηση. Αυτές περιλαμβάνουν την:

- εφαρμογή τεχνικών επεξεργασίας φυσικής γλώσσας για τον προσδιορισμό των ονοματικών οντοτήτων που κατονομάζονται στα νομικά κείμενα και την χρονική διαχείριση των νομικών πόρων π.χ., προσδιορισμός της χρονικής ισχύος ενός νομικού μπλοκ, διάρκεια ισχύος νομικών παραπομπών.

- εφαρμογή τεχνικών μηχανικής μάθησης για την αυτόματη ταξινόμηση των νομικών πηγών με περιγραφές από το EuroVoc, τον πολυγλωσσικό θησαυρό που χρησιμοποιείται για την επισημείωση των νομικών πηγών της Ε.Ε.

- την αυτόματη κωδικοποίηση της (προτεινόμενης) νομοθεσίας βάσει του πρωτοτύπου και των τροποποιητικών του εγγράφων (soft encoding).

- επέκταση της γλώσσας μοντελοποίησης νομικών πηγών για την κάλυψη δικαστικών αποφάσεων.

Όσον αφορά την μοντελοποίηση του δικαίου σε μορφή σύνθετου δικτύου, η πρόταση μας παρέχει μια πρώτη προσέγγιση για την βελτίωση της αποτελεσματικότητας του νομικού συστήματος και νέες ερευνητικές κατευθύνσεις είναι δυνατό να προκύψουν μέσω αυτής. Αυτές, μεταξύ άλλων, περιλαμβάνουν την:

- εφαρμογή τεχνικών ανίχνευσης κοινοτήτων για να ανακαλύψουμε την τομεακή δομή των νομικών πηγών και να εντοπίσουμε ομάδες νομικών πηγών που μοιράζονται κοινές ιδιότητες και έχουν παρόμοιο ρόλο μέσα στο νομικό σώμα.

- ενσωμάτωση νομικών οντολογιών και τεχνολογιών συνδεδεμένων δεδομένων που θα εμπλουτίσουν περαιτέρω την προστιθέμενη αξία του προτεινόμενου μοντέλου αναπαράστασης.

- αξιοποίηση του δικτύου για την γραφική απεικόνιση του δικαίου. Ένα σύστημα απεικόνισης του Δικτύου Νομοθεσίας μπορεί να βοηθήσει τόσο τους πολίτες όσο και τους νομικούς εμπειρογνώμονες να πλοηγηθούν εύκολα στο σώμα του νόμου, επισημαίνοντας πρότυπα, αποκαλύπτοντας συστάδες και συναφείς συνδέσεις, αποκαλύπτοντας επικαλύψεις και πιθανές συγκρούσεις.

- εφαρμογή τεχνικών (topic modelling) για την εξόρυξη των θεμάτων των νομικών πηγών και τον προσδιορισμό των σημασιολογικών ενώσεων τους με βάση τη δομή του δικτύου.

- εφαρμογή τεχνικών συσταδοποίησης για την ανίχνευση αποκλινόντων νομικών πηγών, ακόμη και σε επίπεδο παραγράφων, εάν οι νομικές πηγές έχουν μοντελοποιηθεί καταλλήλως και οι νομικές συνδέσεις έχουν ανιχνευθεί σε επίπεδο παραγράφων κειμένου.

- περαιτέρω αξιολόγηση της ανθεκτικότητας των υπό-δικτύων χρησιμοποιώντας ένα ευρύτερο φάσμα κριτηρίων για τον προσδιορισμό της σημασίας των αφαιρούμενων κόμβων (π.χ., Betweenness, Hits, PageRank).

- πρόσθετες μελέτες με βάση τα χαρακτηριστικά που αποτυπώσαμε ενδέχεται να μας οδηγήσουν σε ένα πλουσιότερο μοντέλο για την δομή και την εξέλιξη του δικτύου.

Σε θέματα (διαφοροποιημένης) ανάκτησης νομικής πληροφορίας σχεδιάζουμε να:

- μελετήσουμε περαιτέρω την αλληλεπίδραση της σχετικότητας και της ποικιλομορφίας σε ιστορικά νομικά ερωτήματα.

- να ενσωματώσουμε και να αξιολογήσουμε στο προτεινόμενο πλαίσιο διαφοροποίησης και άλλες κατηγορίες αλγορίθμων.

Τέλος, για την διαφοροποιημένη ανάκτηση καταχωρήσεων σε κείμενα διαβουλεύσεων και σε κοινωνικά δίκτυα:

- ευελπιστούμε σε αύξηση των δεδομένων σε τυποποιημένη μορφή και με καθορισμένους τρόπους ανάκτησης προκειμένου να αξιολογήσουμε την μεθοδολογία μας σε πραγματικά κείμενα διαβουλεύσεων.

- επιθυμούμε να εξετάσουμε αν υβριδικοί συνδυασμοί αλγορίθμων μπορούν να δώσουν καλύτερα αποτελέσματα.

Βιβλιογραφία

[1] Tommaso Agnoloni, Enrico Francesconi, and Pierluigi Spinosa. xmlegeseditor: an opensource visual xml editor for supporting legal national standards. In *Proceedings of the V legislative XML workshop*, pages 239–251, 2007.

[2] Tommaso Agnoloni and Ugo Pagallo. The case law of the italian constitutional court, its power laws, and the web of scholarly opinions. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15, pages 151–155. ACM Press, 2015. doi:10.1145/2746090.2746108.

[3] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining ,WSDM '09*, pages 5–14. ACM, 2009. doi:10.1145/1498759.1498766.

[4] Elif Aktolga, Irene Ros, and Yannick Assogba. Detecting outlier sections in us congressional legislation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information – SIGIR '11*, pages 235–244. ACM, 2011. doi:10.1145/2009916.2009951.

[5] Réka Albert and A-L Albert László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002. doi:10.1103/RevModPhys.74.47.

[6] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Attack and error tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

[7] Kelli A Alces. Legal Diversification. *Columbia Law Review*, pages 1977–2038, 2013.

[8] Sihem Amer-Yahia and Mounia Lalmas. Xml search: Languages, INEX and scoring. *SIGMOD Rec.*, 35(4):16–23, 2006. doi:10.1145/1228268.1228271.

[9] Anthony B Atkinson. On the measurement of inequality. *Journal of economic theory*, 2(3):244–263, 1970. doi:10.1016/0022-0531(70)90039-6.

[10] Lorenzo Bacci, Pierluigi Spinosa, Carlo Marchetti, and Roberto Battistoni. Automatic mark-up of legislative documents and its application to parallel text genera-

tion. In *Proceedings of LOAIT Workshop*, pages 45–54, 2009. [http://ceur-ws.org/Vol-465/paper6.pdf].

[11] Yaneer Bar-Yam. *Dynamics of complex systems*, volume 213. Westview Press, 1997.

[12] A.L L. Barabási, H Jeong, Z Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, 2002. doi:10.1016/s0378-4371(02)00736-7.

[13] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999. doi:10.1126/science.286.5439.509.

[14] Gioele Barabucci, Luca Cervone, Monica Palmirani, Silvio Peroni, and Fabio Vitali. Multi-layer markup and ontological structures in akoma ntoso. In *AI Approaches to the Complexity of Legal Systems. Complex Systems, the Semantic Web, Ontologies, Argumentation, and Dialogue*, pages 133–149. Springer, 2010. doi:10.1007/978-3-642-16524-5_9.

[15] Danielle Smith Bassett and ED Bullmore. Small-world brain networks. *The neuroscientist*, 12(6):512–523, 2006. doi:10.1177/1073858406293182.

[16] V Richard Benjamins, Pompeu Casanovas, Joost Breuker, and Aldo Gangemi. Law and the semantic web, an introduction. In V. Richard Benjamins, Pompeu Casanovas, Joost Breuker, and Aldo Gangemi, editors, *Law and the Semantic Web*, pages 1–17. Springer, 2005. doi:10.1007/b106624.

[17] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001. doi:10.1038/scientificamerican0501-34.

[18] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria. Automatic semantics extraction in law documents. In *Proceedings of the 10th international conference on Artificial intelligence and law - ICAIL '05*, pages 133–140. ACM Press, 2005. doi:10.1145/1165485.1165506.

[19] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[20] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4–5):175–308, 2006. doi:10.1016/j.physrep.2005.10.009.

[21] Guido Boella, Llio Humphreys, Marco Martin, Piercarlo Rossi, and Leendert van der Torre. Eunomos, a legal document and knowledge management system to build legal services. In Monica Palmirani, Ugo Pagallo, Pompeu Casanovas, and Giovanni Sartor, editors, *International Workshop on AI Approaches to the Complexity of Legal Systems*, pages 131–146. Springer, 2011. doi:10.1007/978-3-642-35731-2.

[22] Alexander Boer, Radboud Winkels, and Fabio Vitali. Metalex xml and the legal knowledge interchange format. In *Computable models of the law*, pages 21–41. Springer, 2008. doi:10.1007/978-3-540-85569-9_2.

[23] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000. doi:10.1016/s1389-1286(00)00083-9.

[24] Michael Busch, Krishna Gade, Brian Larson, Patrick Lok, Samuel Luckenbill, and Jimmy Lin. Earlybird: Real-Time Search at Twitter. In *Proceedings of the 28th International Conference on Data Engineering – ICDE '12*, pages 1360–1369, 2012. doi:10.1109/ICDE.2012.149.

[25] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval – SIGIR '98*, pages 335–336. ACM, 1998. doi:10.1145/290941.291025.

[26] John Carroll, Ted Briscoe, and Antonio Sanfilippo. Parser evaluation: a survey and a new proposal. In *Proceedings of the 1st International Conference on Language Resources and Evaluation – LREC '98*, pages 447–454, 1998.

[27] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*, pages 621–630. ACM, 2009. doi:10.1145/1645953.1646033.

[28] Harr Chen and David R. Karger. Less is more. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval – SIGIR '06*, pages 429–436. ACM, 2006. doi:10.1145/1148170.1148245.

[29] Shiwen Cheng, Anastasios Arvanitis, Marek Chrobak, and Vagelis Hristidis. Multi-Query Diversification in Microblogging Posts. In *Proceedings of the 17th International Conference on Extending Database Technology – EDBT '14*, pages 133–144, 2014. doi:10.5441/002/edbt.2014.13.

[30] Charles L A Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international conference on Research and development in information retrieval – SIGIR '08*, pages 659–666. ACM, 2008. doi:10.1145/1390334.1390446.

[31] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009. doi:10.1137/070710111.

[32] Reuven Cohen, Keren Erez, Daniel Ben-Avraham, and Shlomo Havlin. Breakdown of the internet under intentional attack. *Physical review letters*, 86(16):3682, 2001. doi:10.1103/physrevlett.86.3682.

[33] Steve Cronen-Townsend and W Bruce Croft. Quantifying query ambiguity. In *Proceedings of the second international conference on Human Language Technology Research - HLT '02*. Association for Computational Linguistics (ACL), 2002. doi:10.3115/1289189.1289266.

[34] Paolo Crucitti, Vito Latora, and Massimo Marchiori. Model for cascading failures in complex networks. *Physical Review E*, 69(4):045104, 2004. doi:10.1103/physreve.69.045104.

[35] Emile de Maat, Radboud Winkels, and Tom van Engers. Automated Detection of Reference Structures in Law. In *Proceedings of the 2006 Conference on Legal Knowledge and Information Systems: JURIX 2006: The Nineteenth Annual Conference*, pages 41–50. IOS Press, 2006.

[36] Nicholas Diakopoulos and Mor Naaman. Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work – CSCW '11*, pages 133–142, 2011. doi:10.1145/1958824.1958844.

[37] Marina Drosou and Evaggelia Pitoura. Search result diversification. *ACM SIGMOD Record*, 39(1):41, 2010. doi:10.1145/1860702.1860709.

[38] Paul Erdos and Alfréd Rényi. On the evolution of random graphs. *Bull. Inst. Internat. Statist*, 38:343–347, 1961.

[39] Günes Erkan, Dragomir R Radev, Erkan G., and Dragomir R Radev. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479, 2004.

[40] Erhan Erkut. The discrete p-dispersion problem. *European Journal of Operational Research*, 46(1):48–60, 1990. doi:10.1016/0377-2217(90)90297-O.

[41] Eric Evans. *Domain-driven design: tackling complexity in the heart of software*. Addison-Wesley, 2004.

[42] Atefeh Farzindar and Guy Lapalme. Legal text summarization by exploration of the thematic structures and argumentative roles. In *Text Summarization Branches Out Workshop held in conjunction with ACL*, pages 27–34, 2004. [http://www.aclweb.org/website/old_anthology/W/W04/W04-1006.pdf].

[43] Atefeh Farzindar and Guy Lapalme. Letsum, an automatic legal text summarizing system. In *Legal Knowledge and Information Systems, Jurix 2004: the Sevententh Annual Conference*, pages 11–18. IOS Press, 2004.

[44] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, 2005. doi:10.3115/1219840.1219885.

[45] Bryan Ford. Parsing expression grammars: a recognition-based syntactic foundation. In *Proceedings of the 31st ACM SIGPLAN-SIGACT symposium on Principles of programming languages - POPL '04*, volume 39, pages 111–122. ACM Press, 2004. doi:10.1145/964001.964011.

[46] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010. doi:10.1016/j.physrep.2009.11.002.

[47] J. H. Fowler, T. R. Johnson, J. F. Spriggs, S. Jeon, and P. J. Wahlbeck. Network analysis and the law: Measuring the legal importance of precedents at the u.s. supreme court. *Political Analysis*, 15(3):324–346, 2006. doi:10.1093/pan/mpm011.

[48] James H. Fowler and Sangick Jeon. The authority of supreme court precedent. *Social Networks*, 30(1):16–30, 2008. doi:10.1016/j.socnet.2007.05.001.

[49] Martin Fowler. *Domain Specific Languages*. Addison-Wesley Professional, 2010.

[50] Matias Frosterus, Jouni Tuominen, and Eero Hyvönen. Facilitating re-use of legal data in applications - finnish law as a linked open data service. In *Proceedings of the 20th Legal Knowledge and Information Systems*, JURIX '14, pages 115–124, 2014. doi:10.3233/978-1-61499-468-8-115.

[51] Filippo Galgani, Paul Compton, and Achim Hoffmann. Citation based summarisation of legal texts. In *Proceedings of PRICAI 2012: Trends in Artificial Intelligence: 12th Pacific Rim International Conference on Artificial Intelligence*, pages 40–52. Springer, 2012. doi:10.1007/978-3-642-32695-0_6.

[52] Filippo Galgani, Paul Compton, and Achim Hoffmann. Combining Different Summarization Techniques for Legal Text. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data – HYBRID '12*, pages 115–123. Association for Computational Linguistics, 2012. [http://www.aclweb.org/website/W/W12/W12-05.pdf].

[53] Aldo Gangemi, Maria-Teresa Sagri, and Daniela Tiscornia. Metadata for content description in legal information. In *Proceedings of LegOnt Workshop on Legal Ontologies*, 2003. [http://www.lri.jur.uva.nl/~winkels/LegOnt2003/Gangemi.pdf].

[54] Giorgos Giannopoulos, Marios Koniaris, Ingmar Weber, Alejandro Jaimes, and Timos Sellis. Algorithms and criteria for diversification of news article comments. *Journal of Intelligent Information Systems*, 44(1):1–47, 2015. doi:10.1007/s10844-014-0328-1.

[55] Natalie Glance Gilad Mishne. Leave a Reply: An Analysis of Weblog Comments. In *Proceedings of the Third annual workshop on the Weblogging ecosystem*, 2006. [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.5982].

[56] Colin S. Gillespie. Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software*, 64(2):1–16, 2015. doi:10.18637/jss.v064.i02.

[57] Corrado Gini. Measurement of inequality of incomes. *The Economic Journal*, 31(121):124–126, 1921. doi:10.2307/2223319.

[58] Sreenivas Gollapudi and Aneesh Sharma. An Axiomatic Approach for Result Diversification. In *Proceedings of the 18th international conference on World wide web – WWW '09*, pages 381–390. ACM, 2009. doi:10.1145/1526709.1526761.

[59] Matthias Grabmair, Kevin D. Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R. Walker. Introducing luima: An experiment in legal conceptual retrieval of vaccine injury decisions using a uima type system and tools. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law – ICAIL '15*, pages 69–78. ACM, 2015. doi:10.1145/2746090.2746096.

[60] David J Hand, Heikki Mannila, and Padhraic Smyth. *Principles of data mining*. MIT press, 2001.

[61] S.C. Herring, I. Kouper, J.C. Paolillo, L.A. Scheidt, M. Tyworth, P. Welsch, E. Wright, and Ning Yu Ning Yu. Conversations in the Blogosphere: An Analysis "From the Bottom Up". In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences – HICSS '05*, page 107.2, 2005. doi:10.1109/HICSS.2005.167.

[62] Rinke Hoekstra. The metalex document server. In *Proceedings of the 10th International Semantic Web Conference - ISWC '11*, pages 128–143. Springer, 2011. doi:10.1007/978-3-642-25093-4_9.

[63] Rinke Hoekstra, Joost Breuker, Marcello Di Bello, Alexander Boer, Marcello Di Bello, and Alexander Boer. The LKIF Core ontology of basic legal concepts. In *Proceedings of workshop on Legal Ontologies and Artificial Intelligence Techniques – LOAIT '07*, volume 321, pages 43–63, 2007. [http://ceur-ws.org/Vol-321/LOAIT07-Proceedings.pdf].

[64] Meishan Hu, Aixin Sun, and Ee-Peng Lim. Comments-oriented document summarization: understanding documents with readers' feedback. In *Proceedings of the 31st annual international conference on Research and development in information retrieval – SIGIR '08*, pages 291–298, 2008. doi:10.1145/1390334.1390385.

[65] Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen. Search result diversification based on hierarchical intents. In *Proceedings of the 24th ACM*

International on Conference on Information and Knowledge Management - CIKM '15, pages 63–72, 2015. doi:10.1145/2806416.2806455.

[66] Hirokazu Igari, Akira Shimazu, and Koichiro Ochimizu. Document structure analysis with syntactic model and parsers: Application to legal judgments. In *JSAI International Symposium on A.I.*, pages 126–140. Springer Berlin Heidelberg, 2011. doi:10.1007/978-3-642-32090-3_12.

[67] Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem. Uprising microblogs. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing – SAC '12*, pages 943–948, 2012. doi:10.1145/2245276.2245459.

[68] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000. doi:10.1038/35036627.

[69] Mikko Kivelä, Alexandre Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer Networks. *Journal of Complex Networks*, 2(3):203–271, 2014. doi:10.1093/comnet/cnu016.

[70] Michel Klein, Wouter Van Steenbergen, E.Uijttenbroek, Arno Lodder, and Frank van Harmelen. Thesaurus-based Retrieval of Case Law. In *Proceedings of the 19th International JURIX conference*, 2006.

[71] Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. Legislation as a complex network: Modelling and analysis of European Union legal sources. In *Proceedings of the Twenty Seventh annual conference on Legal Knowledge and Information Systems, (JURIX '14)*, pages 143–152. IOS Press, 2014. doi:10.3233/978-1-61499-468-8-143.

[72] Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. Diversifying the legal order. In Lazaros Iliadis and Ilias Maglogiannis, editors, *IFIP Advances in Information and Communication Technology*, pages 499–509. Springer Nature, 2016. doi:10.1007/978-3-319-44944-9_44.

[73] Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. Multi-dimension diversification in legal information retrieval. In *Proceedings of the 17th International Web Information Systems Engineering – WISE '16*, pages 174–189. Springer, 2016. doi:10.1007/978-3-319-48740-3_12.

[74] Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. Evaluation of diversification techniques for legal information retrieval. *Algorithms*, 10(1):22, 2017. doi:10.3390/a10010022.

[75] Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. Network analysis in the legal domain: A complex model for european union legal sources. *Journal of Complex Networks (Accepted for publication)*, 2017. doi:10.1093/comnet/cnx029.

[76] Marios Koniaris, Giorgos Giannopoulos, Timos Sellis, and Yiannis Vasileiou. Diversifying Microblog Posts. In *Proceedings of the 15th International Web Information Systems Engineering - WISE '14*, pages 189–198. Springer International Publishing, 2014. doi:10.1007/978-3-319-11746-1_14.

[77] Marios Koniaris, George Papastefanatos, Marios Meimaris, and George Alexiou. Introducing solon: A semantic platform for managing legal content. In *Proceedings of the 21st International Conference on Theory and Practice of Digital Libraries - TPDL' 17*, pages 4 (Demo paper – to appear). Springer, 2017.

[78] Marios Koniaris, George Papastefanatos, and Yannis Vassiliou. Towards automatic structuring and semantic indexing of legal documents. In *Proceedings of the 20th Pan-Hellenic Conference on Informatics, PCI '16*. ACM Press, 2016. doi:10.1145/3003733.3003801.

[79] Onur Kucuktunc, B. Barla Cambazoglu, Ingmar Weber, and Hakan Ferhatosmanoglu. A large-scale sentiment analysis for Yahoo! answers. In *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*, pages 633–642. ACM Press, 2012. doi:10.1145/2124295.2124371.

[80] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. *Computer networks*, 31(11):1481–1493, 1999. doi:10.1016/s1389-1286(99)00040-7.

[81] Carl Lagoze, Sandy Payette, Edwin Shin, and Chris Wilper. Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries*, 6(2):124–138, 2006. doi:10.1007/s00799-005-0130-3.

[82] Amy N. Langville and Carl D. Meyer. A survey of eigenvector methods for web information retrieval. *SIAM Review*, 47(1):135–161, 2005. doi:10.1137/s0036144503424786.

[83] Vito Latora and Massimo Marchiori. Efficient behavior of small-world networks. *Physical review letters*, 87(19):198701, 2001. doi:10.1103/PhysRevLett.87.198701.

[84] Philip Leith. The rise and fall of the legal expert system. *European Journal of Law and Technology*, 1(1):94–106, 2010. doi:10.1080/13600869.2016.1232465.

[85] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007. doi:10.1145/1217299.1217301.

[86] Nicola Lettieri, Antonio Altamura, Armando Faggiano, and Delfina Malandrino. A computational approach for the experimental study of eu case law: analysis and

implementation. *Social Network Analysis and Mining*, 6(1):1–17, 2016. doi:10.1007/s13278-016-0365-6.

[87] Qing Li, Jia Wang, Yuanzhu Peter Chen, and Zhangxi Lin. User comments for news recommendation in forum-based social media. *Information Sciences: an International Journal*, 180(24):4929–4939, 2010. doi:10.1016/j.ins.2010.08.044.

[88] Ciciliati F. Lima JAO. LexML Brasil, Parte 3 - LexML XML Schema, version 1.0. Technical report, 2008. [http://projeto.lexml.gov.br/documentacao/Parte-3-XML-Schema.pdf].

[89] Qiang Lu, Jack G. Conrad, Khalid Al-Kofahi, and William Keenan. Legal document clustering with built-in topic segmentation. In *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, pages 383–392. ACM Press, 2011. doi:10.1145/2063576.2063636.

[90] Caterina Lupo, Fabio Vitali, Enrico Francesconi, Monica Palmirani, Radboud Winkels, Emile de Maat, Alexander Boer, and Paolo Mascellani. ESTRELLA Project, Deliverable D3.1 - General XML format(s) for legal Sources, version 1.0. Technical report, 2007. [http://www.estrellaproject.org/doc/D3.1-General-XML-formats-For-Legal-Sources.pdf].

[91] Yonatan Lupu and Erik Voeten. Precedent in international courts: A network analysis of case citations by the european court of human rights. *British Journal of Political Science*, 42(02):413–439, 2012. doi:10.1017/s0007123411000433.

[92] Christos Makris, Yannis Plegas, Yannis C Stamatiou, Elias C Stavropoulos, and Athanasios K Tsakalidis. Reducing redundant information in search results employing approximation algorithms. In *International Conference on Database and Expert Systems Applications*, pages 240–247. Springer Science Business Media, 2014. doi:10.1007/978-3-319-10085-2_22.

[93] Andrea Marchetti, Fabrizio Megale, Enrico Seta, and Fabio Vitali. Using xml as a means to access legislative documents: Italian and foreign experiences. *ACM SIGAPP Applied Computing Review*, 10(1):54–62, 2002. doi:10.1145/568235.568246.

[94] Pierre Mazzega, Danièle Bourcier, and Romain Boulet. The network of french legal codes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law - ICAIL '09*, pages 236–237. ACM Press, 2009. doi:10.1145/1568234.1568271.

[95] Qiaozhu Mei, Jian Guo, and Dragomir Radev. DivRank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining – KDD '10*, pages 1009–1018. ACM Press, 2010. doi:10.1145/1835804.1835931.

[96] Marios Meimaris, George Alexiou, and George Papastefanatos. Linkzoo: A linked data platform for collaborative management of heterogeneous resources. In *The Semantic Web: ESWC 2014 Satellite Events*, pages 407–412. Springer, 2014. doi:10.1007/978-3-319-11955-7_57.

[97] Eneldo Loza Mencía and Johannes Fürnkranz. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *Machine Learning and Knowledge Discovery in Databases*, pages 50–65. Springer Nature, 2008. doi:10.1007/978-3-540-87481-2_4.

[98] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967. doi:10.1037/e400002009-005.

[99] Marie-Francine Moens. Innovative techniques for legal text retrieval. *Artificial Intelligence and Law*, 9(1):29–57, 2001. doi:10.1023/A:1011297104922.

[100] Marie-Francine Moens. Summarizing court decisions. *Information Processing & Management*, 43(6):1748–1764, 2007. doi:10.1016/j.ipm.2007.01.005.

[101] Jose M Montoya and Ricard V Solé. Small world patterns in food webs. *Journal of theoretical biology*, 214(3):405–412, 2002. doi:10.1006/jtbi.2001.2460.

[102] Sean A. Munson and Paul Resnick. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '10*, pages 1457–1466, 2010. doi:10.1145/1753326.1753543.

[103] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003. doi:10.1137/S003614450342480.

[104] Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003. doi:10.1103/physreve.67.026126.

[105] Marc van Opijnen, Nico Verwer, and Jan Meijer. Beyond the experiment: the extendable legal link extractor. Workshop on automated detection, extraction and analysis of semantic information in legal texts, 2015. [https://ssrn.com/abstract=2626521].

[106] Jahna Otterbacher, Gunes Erkan, and Dragomir R. Radev. Biased LexRank: Passage retrieval using random walks with question-based priors. *Information Processing & Management*, 45(1):42–54, jan 2009. doi:10.1016/j.ipm.2008.06.004.

[107] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. [http://ilpubs.stanford.edu:8090/422/].

[108] Abdul Paliwala. A history of legal informatics. *European Journal of Law and Technology*, 1(1), 2010.

[109] Monica Palmirani and Raffaella Brighi. An xml editor for legal information management. In *Proceedings of Electronic Government: Second International Conference – EGOV '03*, pages 421–429. Springer, 2003. doi:10.1007/10929179_76.

[110] Souneil Park, Minsam Ko, Jungwoo Kim, Ying Liu, and Junehwa Song. The politics of comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work – CSCW '11*, pages 113–122, 2011. doi:10.1145/1958824.1958842.

[111] Terence Parr. *Language Implementation Patterns: Create Your Own Domain-Specific and General Programming Languages*. Pragmatic Bookshelf, 1st edition, 2009.

[112] Terence Parr, Sam Harwell, and Kathleen Fisher. Adaptive ll (*) parsing: the power of dynamic analysis. *ACM SIGPLAN Notices*, 49(10):579–598, 2014. doi:10.1145/2714064.2660202.

[113] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001. doi:10.1103/physrevlett.86.3200.

[114] Martin Potthast. Measuring the descriptiveness of web comments. In *Proceedings of the 32nd international conference on Research and development in information retrieval – SIGIR '09*, pages 724–725, 2009. doi:10.1145/1571941.1572097.

[115] Filip Radlinski, Paul N Bennett, Ben Carterette, and Thorsten Joachims. Redundancy diversity and interdependent document relevance. *ACM SIGIR Forum*, 43(2):46, 2009. doi:10.1145/1670564.1670572.

[116] Karthik Raman, Pannaga Shivaswamy, and Thorsten Joachims. Online learning to diversify from implicit feedback. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD' 12*, pages 705–713. ACM, 2012. doi:10.1145/2339530.2339642.

[117] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2), 2003. doi:10.1103/physreve.67.026112.

[118] Gerasimos Razis and Ioannis Anagnostopoulos. InfluenceTracker: Rating the impact of a Twitter account. In *Proceedings Artificial Intelligence Applications and Innovations - (AIAI '14) -Workshops: CoPA, MHDW, IIVC, and MT4BD*, pages 184 –195. Springer Berlin Heidelberg, 2014. doi:10.1007/978-3-662-44722-2_20.

[119] Independent Authority for Public Revenue. Business plan, 2016. In Greek [http://www.publicrevenue.gr/kpi/static/doc/epixirisiako_sxedio_ggde_2016_v5.pdf].

[120] Inter-Parliamentary Union. World e-Parliament Report. 2016. [http://www.ipu.org/pdf/publications/eparl16-en.pdf].

[121] Zhaochun Ren, Shangsong Liang, Edgar Meij, and Maarten de Rijke. Personalized time-aware tweets summarization. In *Proceedings of the 36th international conference on Research and development in information retrieval – SIGIR '13*, pages 513–522, 2013. doi:10.1145/2484028.2484052.

[122] Jesus A. Rodriguez Perez, Yashar Moshfeghi, and Joemon M. Jose. On using inter-document relations in microblog retrieval. In *Companion Proceedings of the 22nd international conference on World Wide Web companion – WWW '13*, pages 75–76, 2013. doi:10.1.1.402.2687.

[123] Martin Rosvall and Carl T. Bergstrom. Mapping change in large networks. *PLoS ONE*, 5(1):1–7, 2010. doi:10.1371/journal.pone.0008694.

[124] JB Ruhl and Daniel Martin Katz. Measuring, monitoring, and managing legal complexity. *Iowa Law Review*, 101, 2015. [https://ssrn.com/abstract=2566535].

[125] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975. doi:10.1145/361219.361220.

[126] Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010. doi:10.1561/1500000009.

[127] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. Search result diversification. *Foundations and Trends in Information Retrieval*, 9(1):1–90, 2015. URL: 10.1561/1500000040, doi:10.1561/1500000040.

[128] Rodrygo L.T. T Santos, Craig Macdonald, and Iadh Ounis. Exploiting Query Reformulations for Web Search Result Diversification. pages 881–890, 2010. doi:10.1145/1772690.1772780.

[129] M. Saravanan, B. Ravindran, and S. Raman. Improving legal information retrieval using an ontological framework. *Artif Intell Law*, 17(2):101–124, 2009. doi:10.1007/s10506-009-9075-y.

[130] Erich Schweighofer and Doris Liebwald. Advanced lexical ontologies and hybrid knowledge based systems: First steps to a dynamic legal electronic commentary. *Artificial Intelligence and Law*, 15(2):103–115, 2007. doi:10.1007/s10506-007-9029-1.

[131] Parongama Sen, Subinay Dasgupta, Arnab Chatterjee, PA Sreeram, G Mukherjee, and SS Manna. Small-world properties of the indian railway network. *Physical Review E*, 67(3):036106, mar 2003. doi:10.1103/physreve.67.036106.

[132] Erez Shmueli, Amit Kagian, Yehuda Koren, and Ronny Lempel. Care to comment?: recommendations for commenting on news stories. In *Proceedings of the 21st international conference on World Wide Web – WWW '12*, pages 429–438, 2012. doi:10.1145/2187836.2187895.

[133] Herbert A Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955. doi:10.2307/2333389.

[134] Jaspreet Singh, Wolfgang Nejdl, and Avishek Anand. History by diversity: Helping historians search news archives. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval – CHIIR '16*, pages 183–192. ACM, 2016. doi:10.1145/2854946.2854959.

[135] Thomas A.C. Smith. The web of law. *San Diego Legal Studies Research Paper*, 6(11):1–39, 2005. doi:10.2139/ssrn.642863.

[136] Kai Song, Yonghong Tian, Wen Gao, and Tiejun Huang. Diversifying the image retrieval results. In *Proceedings of the 14th annual ACM international conference on Multimedia - MULTIMEDIA '06*, pages 707–710. ACM, 2006. doi:10.1145/1180639.1180789.

[137] Fabien Tarissan and Raphaëlle Nollez-Goldbach. Analysing the first case of the international criminal court from a network-science perspective. *Journal of Complex Networks*, pages 616–634, 2016. doi:10.1093/comnet/cnw002.

[138] Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. #TwitterSearch: A Comparison of Microblog Search and Web Search. In *Proceedings of the fourth ACM international conference on Web search and data mining – WSDM '11*, pages 35–44. ACM Press, 2011. doi:10.1145/1935826.1935842.

[139] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment in short strength detection informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010. doi:10.1002/asi.21416.

[140] Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke. Predicting the volume of comments on online news stories. In *Proceeding of the 18th conference on Information and knowledge management – CIKM '09*, pages 1765–1768, 2009. doi:10.1145/1645953.1646225.

[141] Manos Tsagkias, Wouter Weerkamp, and Maarten De Rijke. News comments: Exploring, modeling, and online prediction. In *Proceedings of the 32nd European Conference on Information Retrieval – ECIR '10*, pages 191–203, 2010. doi:10.1007/978-3-642-12275-0-19.

[142] Stanislav Bederev v Ireland. The attorney general and the director of public prosecutions. *Irish Court of Appeal*, 1409, 2014.

[143] Saskia Van De Ven, Rinke Hoekstra, Radboud Winkels, Emile de Maat, and Ádám Kollár. Metavex: Regulation drafting meets the semantic web. In *Computable Models of the Law*, pages 42–55. Springer, 2008. doi:10.1007/978-3-540-85569-9_3.

[144] Marc van Opijnen. Citation analysis and beyond: in search of indicators measuring case law importance. In *Proceedings of the Twenty Fifth annual conference on Legal Knowledge and Information Systems - JURIX '12*, volume 250 of *Frontiers in Artificial Intelligence and Applications*, pages 95–104. IOS Press, 2012. doi:10.3233/978-1-61499-167-0-95.

[145] Marc van Opijnen. A model for automated rating of case law. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law – ICAIL '13*, pages 140–149. ACM Press, 2013. doi:10.1145/2514601.2514617.

[146] Marc van Opijnen and Cristiana Santos. On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law*, 25(1):65–87, 2017. doi:10.1007/s10506-017-9195-8.

[147] Erik Vee, Utkarsh Srivastava, Jayavel Shanmugasundaram, Prashant Bhat, and Sihem Amer Yahia. Efficient computation of diverse query results. In *Proceedings of the 24th International Conference on Data Engineering – ICDE '08*, pages 228–236, 2008. doi:10.1109/ICDE.2008.4497431.

[148] Andreas Wagner and David A Fell. The small world inside large metabolic networks. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1478):1803–1810, 2001. doi:10.1098/rspb.2001.1711.

[149] Xiao Fan Wang and Guanrong Chen. Complex networks: small-world, scale-free and beyond. *IEEE circuits and systems magazine*, 3(1):6–20, 2003. doi:10.1109/mcas.2003.1228503.

[150] D J Watts and S H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998. doi:10.1038/30918.

[151] Frank Webster. *Theories of the information society*. Routledge, 2014.

[152] R Winkels, J Ruyter, and H Kroese. Determining authority of dutch case law. In *Legal Knowledge and Information Systems - JURIX 2011: The Twenty-Fourth International Conference*, volume 235, pages 103–112, 2011.

[153] Armin Wittfoth, Philip Chung, Graham Greenleaf, and Andrew Mowbray. Austlii's point-in-time legislation system: A generic pit system for presenting legislation. Technical report, 2005. [http://portsea.austlii.edu.au/pit/papers/PiT_background_2005.rtf].

[154] David Wong, Siamak Faridani, Ephrat Bitton, Björn Hartmann, and Ken Goldberg. The diversity donut: enabling participant control over the diversity of recommended responses. In *Proceedings of the 2011 annual conference on Human factors in computing systems extended abstracts – CHI EA '11*, pages 1471–1476. ACM Press, 2011. doi:10.1145/1979742.1979793.

[155] Cheng Xiang Zhai, William W Cohen, and John Lafferty. Beyond independent relevance. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval – SIGIR '03*. ACM, 2003. doi:10.1145/860435.860440.

[156] Benyu Zhang, Hua Li, Yi Liu, Lei Ji, Wensi Xi, Weiguo Fan, Zheng Chen, and Wei-Ying Ma. Improving web search results using affinity graph. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval – SIGIR '05*, pages 504–511. ACM, 2005. doi:10.1145/1076034.1076120.

[157] Paul Zhang and Lavanya Koppaka. Semantics-based legal citation network. In *Proceedings of the 11th international conference on Artificial intelligence and law – ICAIL '07*, pages 123–130. ACM Press, 2007. doi:10.1145/1276318.1276342.

[158] Xin Zhang, Ben He, Tiejian Luo, and Baobin Li. Query-biased learning to rank for real-time twitter search. In *Proceedings of the 21st international conference on Information and knowledge management – CIKM '12*, pages 1915–1919, 2012. doi:10.1145/2396761.2398543.

[159] Xiaojin Zhu, Andrew B. Goldberg, Jurgen Van Gael, David Andrzejewski, Jurgen Van Gael, and David Andrzejewski. Improving Diversity in Ranking using Absorbing Random Walks. In *Proceedings of the Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics – NAACL-HLT '07*, pages 97–104, 2007. [http://www.aclweb.org/website/old_anthology/N/N07/N07-1.pdf].

[160] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web - WWW '05*, pages 22–32. ACM, 2005. doi:10.1145/1060745.1060754.

Παραρτημα Α΄

Συγκεντρωτικά αποτελέσματα των υπό αξιολόγηση μεθόδων διαφοροποίησης

Πίνακας Α′.1: Επίδοση των αξιολογούμενων μεθόδων για τιμές παραμέτρων $|N| = 100$, $k = 30$ και trade-off $\lambda \in [0..1]$ αυξανόμενη με βήμα 0.1. Οι υψηλότερες βαθμολογίες εμφανίζονται με έντονους χαρακτήρες. Στατιστικώς σημαντικές τιμές, χρησιμοποιώντας το paired two-sided t-test, με τιμές $p_{value} < 0.05$ σημειώνονται με $^\circ$ και για $p_{value} < 0.01$ με *.

	a-NDCG				nERR-IA				ST Recall			
Αλγόριθμος	@5	@10	@20	@30	@5	@10	@20	@30	@5	@10	@20	@30
$\lambda = 0.1$												
baseline	0.5044	0.5498	0.6028	0.6292	0.4925	0.5153	0.5333	0.5395	0.5827	0.7260	0.8464	0.9010
MMR	0.5187	**0.5785***	**0.642***	**0.6676***	**0.5041**	**0.5341**	**0.5559**$^\circ$	**0.562**$^\circ$	0.6145°	**0.7875***	**0.9135***	**0.9543***
Max-sum	0.5170	0.5699°	0.6276*	0.6541*	0.5022	0.5290	0.5486	0.5549	0.6083	0.7626°	0.8851*	0.9294*
Max-min	0.5188	0.5749*	0.6365*	0.6633*	0.5029	0.5313	0.5526°	0.5589°	0.6173°	0.7820*	0.8990*	0.9481*
Mono-objective	0.5052	0.5584	0.6160	0.6450	0.4919	0.5184	0.5382	0.5450	0.5889	0.7543	0.8740°	0.9273*
LexRank	0.4152*	0.4357*	0.4823*	0.5154*	0.4160*	0.4258*	0.4413*	0.4491*	0.4228*	0.5329*	0.6713*	0.7647*
Biased LexRank	0.4155*	0.4373*	0.4833*	0.5160*	0.4163*	0.4268*	0.4421*	0.4498*	0.4228*	0.5370*	0.6734*	0.7654*
DivRank	**0.5195**	0.5774*	0.6304*	0.6543*	0.5035	0.5328	0.5511	0.5567	**0.6208**$^\circ$	0.7820*	0.8976*	0.9384*
Grasshopper	0.4368*	0.4611*	0.5069*	0.5389*	0.4359*	0.4476*	0.4630*	0.4705*	0.4567*	0.5758*	0.7059*	0.7931*
$\lambda = 0.2$												
baseline	0.5044	0.5498	0.6028	0.6292	0.4925	0.5153	0.5333	0.5395	0.5827	0.7260	0.8464	0.9010
MMR	**0.5356***	**0.6015***	**0.6607***	**0.6845***	**0.5167**$^\circ$	**0.5499***	**0.5704***	**0.576***	**0.6547***	**0.8388***	**0.9322***	**0.9696***
Max-sum	0.5277°	0.5838*	0.6397*	0.6637*	0.5080	0.5366°	0.5557°	0.5613°	0.6422*	0.7993*	0.9017*	0.9398*
Max-min	0.5309*	0.5929*	0.6524*	0.6771*	0.5113	0.5425*	0.5629*	0.5687*	0.6533*	0.8187*	0.9246*	0.9640*
Mono-objective	0.5102	0.5658	0.6284*	0.6550*	0.4947	0.5226	0.5439	0.5502	0.6035	0.7654*	0.8941*	0.9398*
LexRank	0.4152*	0.4357*	0.4823*	0.5154*	0.4160*	0.4258*	0.4413*	0.4491*	0.4228*	0.5329*	0.6713*	0.7647*
Biased LexRank	0.4172*	0.4387*	0.4850*	0.5173*	0.4176*	0.4280*	0.4433*	0.4509*	0.4277*	0.5391*	0.6761*	0.7668*
DivRank	0.5077	0.5657	0.6209°	0.6454°	0.4902	0.5196	0.5386	0.5444	0.6159°	0.7931*	0.9100*	0.9453*
Grasshopper	0.4403*	0.4647*	0.5099*	0.5419*	0.4384*	0.4502*	0.4654*	0.4729*	0.4657*	0.5799*	0.7100*	0.7965*
$\lambda = 0.3$												
baseline	0.5044	0.5498	0.6028	0.6292	0.4925	0.5153	0.5333	0.5395	0.5827	0.7260	0.8464	0.9010
MMR	**0.547***	**0.6142***	**0.6702***	**0.6912***	**0.5242***	**0.5584***	**0.5778***	**0.5828***	**0.6955***	**0.8581***	**0.9439***	**0.9682***

Συνέχεια στην επόμενη σελίδα ...

Πίνακας Α'.1 – *Συνέχεια από την προηγούμενη σελίδα*

	a-nDCG				nERR-IA				ST Recall			
Αλγόριθμος	@5	@10	@20	@30	@5	@10	@20	@30	@5	@10	@20	@30
Max-sum	0.5308*	0.5911*	0.6473*	0.6708*	0.5109	0.5416*	0.5610*	0.5666*	0.6512*	0.8111*	0.9093*	0.9460*
Max-min	0.5394*	0.6022*	0.6610*	0.6840*	0.5170°	0.5490*	0.5693*	0.5748*	0.6775*	0.8339*	0.9343*	0.9675*
Mono-objective	0.5150	0.5731°	0.6361*	0.6621*	0.4988	0.5280	0.5497	0.5558	0.6159°	0.7779*	0.9059*	0.9481*
LexRank	0.4152*	0.4357*	0.4823*	0.5154*	0.4160*	0.4258*	0.4413*	0.4491*	0.4228*	0.5329*	0.6713*	0.7647*
Biased LexRank	0.4194*	0.4414*	0.4867*	0.5190*	0.4201*	0.4305*	0.4456*	0.4532*	0.4298*	0.5433*	0.6754*	0.7675*
DivRank	0.5261°	0.5779*	0.6316*	0.6566*	0.5109	0.5371°	0.5557°	0.5616°	0.6394*	0.7882*	0.8886*	0.9356*
Grasshopper	0.4421*	0.4667*	0.5113*	0.5430*	0.4395*	0.4514*	0.4664*	0.4738*	0.4713*	0.5848*	0.7121*	0.7965*
$\lambda = 0.4$												
baseline	0.5044	0.5498	0.6028	0.6292	0.4925	0.5153	0.5333	0.5395	0.5827	0.7260	0.8464	0.9010
MMR	**0.5527***	**0.6226***	**0.677***	**0.696***	**0.5291***	**0.5647***	**0.5836***	**0.5881***	**0.7093***	**0.8727***	**0.9481***	0.9696*
Max-sum	0.5425*	0.5996*	0.6580*	0.6798*	0.5214*	0.5505*	0.5708*	0.5759*	0.6740*	0.8187*	0.9183*	0.9522*
Max-min	0.5447*	0.6073*	0.6659*	0.6879*	0.5206*	0.5524*	0.5726*	0.5778*	0.6962*	0.8422*	0.9405*	**0.9723***
Mono-objective	0.5186	0.5802*	0.6422*	0.6671*	0.5025	0.5334	0.5546°	0.5605°	0.6208°	0.7903*	0.9114*	0.9543*
LexRank	0.4152*	0.4357*	0.4823*	0.5154*	0.4160*	0.4258*	0.4413*	0.4491*	0.4228*	0.5329*	0.6713*	0.7647*
Biased LexRank	0.4202*	0.4415*	0.4879*	0.5199*	0.4208*	0.4310*	0.4465*	0.4541*	0.4298*	0.5412*	0.6768*	0.7682*
DivRank	0.5140	0.5710°	0.6298*	0.6540*	0.4968	0.5258	0.5460	0.5517	0.6111	0.7785*	0.9045*	0.9439*
Grasshopper	0.4421*	0.4673*	0.5117*	0.5432*	0.4389*	0.4513*	0.4662*	0.4736*	0.4740*	0.5896*	0.7142*	0.7979*
$\lambda = 0.5$												
baseline	0.5044	0.5498	0.6028	0.6292	0.4925	0.5153	0.5333	0.5395	0.5827	0.7260	0.8464	0.9010
MMR	**0.557***	**0.6278***	**0.6796***	**0.6991***	**0.5329***	**0.5691***	**0.5872***	**0.5918***	**0.7218***	**0.8844***	**0.9495***	**0.9737***
Max-sum	0.5397*	0.6052*	0.6590*	0.6812*	0.5173°	0.5506*	0.5692*	0.5744*	0.6824*	0.8381*	0.9211*	0.9571*
Max-min	0.5477*	0.6130*	0.6701*	0.6913*	0.5233*	0.5567*	0.5764*	0.5814*	0.7024*	0.8554*	0.9433*	**0.9737***
Mono-objective	0.5208	0.5816*	0.6446*	0.6693*	0.5024	0.5331	0.5547°	0.5605°	0.6318*	0.7965*	0.9183*	0.9571*
LexRank	0.4152*	0.4357*	0.4823*	0.5154*	0.4160*	0.4258*	0.4413*	0.4491*	0.4228*	0.5329*	0.6713*	0.7647*
Biased LexRank	0.4203*	0.4429*	0.4887*	0.5204*	0.4209*	0.4319*	0.4472*	0.4547*	0.4291*	0.5446*	0.6782*	0.7675*
DivRank	0.5252°	0.5810*	0.6327*	0.6542*	0.5044	0.5329	0.5507	0.5558	0.6450*	0.8000*	0.8976*	0.9280*

Συνέχεια στην επόμενη σελίδα ...

Πίνακας Α′.1 – *Συνέχεια από την προηγούμενη σελίδα*

	a-nDCG				nERR-IA				ST Recall			
Αλγόριθμος	@5	@10	@20	@30	@5	@10	@20	@30	@5	@10	@20	@30
Grasshopper	0.4402*	0.4654*	0.5103*	0.5423*	0.4371*	0.4494*	0.4645*	0.4720*	0.4713*	0.5869*	0.7128*	0.7986*
$\lambda = 0.6$												
baseline	0.5044	0.5498	0.6028	0.6292	0.4925	0.5153	0.5333	0.5395	0.5827	0.7260	0.8464	0.9010
MMR	**0.5628***	**0.6315***	**0.6825***	**0.7021***	**0.5377***	**0.5729***	**0.5906***	**0.5953***	**0.7363***	**0.8872***	**0.9516***	**0.9744***
Max-sum	0.5482*	0.6103*	0.6656*	0.6861*	0.5250*	0.5566*	0.5760*	0.5809*	0.7024*	0.8422*	0.9260*	0.9557*
Max-min	0.5501*	0.6176*	0.6733*	0.6939*	0.5267*	0.5610*	0.5803*	0.5852*	0.7038*	0.8602*	0.9467*	0.9723*
Mono-objective	0.5196	0.5824*	0.6454*	0.6699*	0.5005	0.5323	0.5540°	0.5598°	0.6325*	0.8014*	0.9218*	0.9606*
LexRank	0.4152*	0.4357*	0.4823*	0.5154*	0.4160*	0.4258*	0.4413*	0.4491*	0.4228*	0.5329*	0.6713*	0.7647*
Biased LexRank	0.4225*	0.4442*	0.4905*	0.5217*	0.4230*	0.4336*	0.4490*	0.4564*	0.4332*	0.5446*	0.6817*	0.7675*
DivRank	0.5185	0.5711°	0.6253*	0.6532*	0.4986	0.5256	0.5442	0.5508	0.6401*	0.7945*	0.9017*	0.9467*
Grasshopper	0.4374*	0.4619*	0.5077*	0.5394*	0.4346*	0.4465*	0.4619*	0.4693*	0.4657*	0.5806*	0.7107*	0.7958*
$\lambda = 0.7$												
baseline	0.5044	0.5498	0.6028	0.6292	0.4925	0.5153	0.5333	0.5395	0.5827	0.7260	0.8464	0.9010
MMR	**0.5662***	**0.6333***	**0.6829***	**0.7026***	**0.5393***	**0.5734***	**0.5907***	**0.5954***	**0.7467***	**0.8893***	**0.9516***	**0.9744***
Max-sum	0.5516*	0.6134*	0.6690*	0.6886*	0.5276*	0.5590*	0.5784*	0.5831*	0.7073*	0.8450*	0.9280*	0.9543*
Max-min	0.5536*	0.6191*	0.6749*	0.6941*	0.5283*	0.5619*	0.5812*	0.5858*	0.7093*	0.8623*	0.9481*	0.9702*
Mono-objective	0.5215	0.5859*	0.6473*	0.6720*	0.5028	0.5355°	0.5567°	0.5626°	0.6353*	0.8083*	0.9190*	0.9599*
LexRank	0.4152*	0.4357*	0.4823*	0.5154*	0.4160*	0.4258*	0.4413*	0.4491*	0.4228*	0.5329*	0.6713*	0.7647*
Biased LexRank	0.4239*	0.4456*	0.4917*	0.5229*	0.4242*	0.4347*	0.4501*	0.4574*	0.4353*	0.5467*	0.6837*	0.7689*
DivRank	0.5123	0.5654	0.6170	0.6446	0.4933	0.5202	0.5381	0.5446	0.6353*	0.7896*	0.8865*	0.9391*
Grasshopper	0.4312*	0.4566*	0.5034*	0.5359*	0.4295*	0.4420*	0.4577*	0.4653*	0.4533*	0.5702*	0.7038*	0.7945*
$\lambda = 0.8$												
baseline	0.5044	0.5498	0.6028	0.6292	0.4925	0.5153	0.5333	0.5395	0.5827	0.7260	0.8464	0.9010
MMR	**0.5676***	**0.6312***	**0.6834***	**0.7024***	**0.5397***	**0.5723***	**0.5906***	**0.5952***	**0.7502***	**0.881***	**0.9516***	**0.9737***
Max-sum	0.5494*	0.6140*	0.6700*	0.6889*	0.5248*	0.5577*	0.5772*	0.5817*	0.7093*	0.8512*	0.9343*	0.9571*

Συνέχεια στην επόμενη σελίδα ...

Πίνακας Α'.1 – *Συνέχεια από την προηγούμενη σελίδα*

	a-nDCG				nERR-IA				ST Recall			
Αλγόριθμος	@5	@10	@20	@30	@5	@10	@20	@30	@5	@10	@20	@30
Max-min	0.5547*	0.6228*	0.6767*	0.6957*	0.5291*	0.5640*	0.5827*	0.5872*	0.7156*	0.8706*	0.9509*	0.9716*
Mono-objective	0.5234	0.5880*	0.6493*	0.6735*	0.5040	0.5369°	0.5580*	0.5636*	0.6443*	0.8118*	0.9232*	0.9626*
LexRank	0.4152*	0.4357*	0.4823*	0.5154*	0.4160*	0.4258*	0.4413*	0.4491*	0.4228*	0.5329*	0.6713*	0.7647*
Biased LexRank	0.4252*	0.4479*	0.4937*	0.5247*	0.4261*	0.4371*	0.4524*	0.4596*	0.4346*	0.5488*	0.6858*	0.7709*
DivRank	0.5155	0.5735*	0.6304*	0.6517*	0.4995	0.5289	0.5484	0.5534	0.6090	0.7806*	0.8976*	0.9280*
Grasshopper	0.4262*	0.4515*	0.4986*	0.5320*	0.4249*	0.4372*	0.4530*	0.4608*	0.4457*	0.5647*	0.6969*	0.7931*
$\lambda = 0.9$												
baseline	0.5044	0.5498	0.6028	0.6292	0.4925	0.5153	0.5333	0.5395	0.5827	0.7260	0.8464	0.9010
MMR	**0.5647***	**0.6306***	**0.6834***	**0.7018***	**0.5381***	**0.5718***	**0.5902***	**0.5946***	**0.7439***	**0.8817***	**0.9529***	**0.9737***
Max-sum	0.5429*	0.6136*	0.6699*	0.6884*	0.5188*	0.5551*	0.5748*	0.5792*	0.6997*	0.8554*	0.9419*	0.9626*
Max-min	0.5570*	0.6244*	0.6781*	0.6971*	0.5311*	0.5657*	0.5844*	0.5889*	0.7211*	0.8727*	0.9495*	0.9716*
Mono-objective	0.5238	0.5894*	0.6502*	0.6734*	0.5037	0.5371°	0.5580*	0.5635*	0.6471*	0.8166*	0.9260*	0.9619*
LexRank	0.4152*	0.4357*	0.4823*	0.5154*	0.4160*	0.4258*	0.4413*	0.4491*	0.4228*	0.5329*	0.6713*	0.7647*
Biased LexRank	0.4250*	0.4486*	0.4948*	0.5254*	0.4262*	0.4377*	0.4532*	0.4604*	0.4325*	0.5502*	0.6872*	0.7702*
DivRank	0.5177	0.5721°	0.6295*	0.6511*	0.5001	0.5275	0.5473	0.5524	0.6187°	0.7785*	0.8969*	0.9280*
Grasshopper	0.4199*	0.4438*	0.4915*	0.5264*	0.4193*	0.4309*	0.4470*	0.4552*	0.4332*	0.5495*	0.6851*	0.7882*

Παραρτημα Β΄

Μεταφράσεις Αγγλικών Όρων

Ελληνικός όρος	Αγγλικός όρος
ακρίβεια	precision
ανάκτηση πληροφορίας	information retrieval (IR)
νομική πληροφορική	legal informatics
μοντέλο διανυσματικού χώρου	vector space model
σχήμα ευρετηρίασης	indexing schema
διανύσματα όρων	term vectors
τυχαίος περίπατος	random walk
χρονικά μεταβαλλόμενος τυχαίος περίπατος	time-variant random walk
συλλογή νομικών εγγράφων	legal corpus
δίκτυα μικρού-κόσμου	small-world
δίκτυα νόμου δύναμης	power-law
δίκτυα ελεύθερης-κλίμακας	scale-free
κανονικά δίκτυα	regular
τυχαία μοντέλα δικτύων	random
κατανομή βαθμών	degree distribution
βαθμός διαμεσολάβησης	betweenness
γιγάντια συνεκτική συνιστώσα	giant connected component
έλεγχος t κατά ζεύγη	paired two-sided t-test

Παραρτημα Γ′

Βιογραφικό Σημείωμα

Στοιχεία Επικοινωνίας

Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων
Τομέας Τεχνολογίας Υπολογιστών και Πληροφορικής
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εθνικό Μετσόβιο Πολυτεχνείο
Ηρώων Πολυτεχνείου 9, Πολυτεχνειούπολη Ζωγράφου
157 80, Αθήνα, Ελλάδα
Τηλέφωνο: +30-210-7721602, +30-210-7721402
Fax: +30-210-7721442
Διεύθυνση ηλ. ταχυδρομείου e-mail: mkoniari@dbnet.ntua.gr
Προσωπική ιστοσελίδα: `http://www.dbnet.ece.ntua.gr/~mkoniari`

Σπουδές

– 1993 - 1999: Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

– 2005 - 2007: Μεταπτυχιακό Δίπλωμα Εξειδίκευσης στη Διοίκηση Επιχειρήσεων (MBA) Ο.Π.Α. - Ε.Μ.Π.

– 2009 - 2012: Μεταπτυχιακό Δίπλωμα Εξειδίκευσης στη Διαχείριση Τεχνικών Έργων Ε.Α.Π.

Δημοσιεύσεις

1. Marios Koniaris George Papastefanatos, Marios Meimaris, and George Alexiou. Introducing solon: A semantic platform for managing legal content. In *Proceedings of the*

21st International Conference on Theory and Practice of Digital Libraries - TPDL' 17, 4 pages (Demo paper – to appear). Springer, 2017.

2. Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. Network analysis in the legal domain: A complex model for european union legal sources. *Journal of Complex Networks (Accepted for publication)*, 2017. URL: `http://dx.doi.org/10.1093/comnet/cnx029`.

3. Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. Evaluation of diversification techniques for legal information retrieval. *Algorithms*, 10(1):22, 2017. `doi:10.3390/a10010022`.

4. Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. Multi-dimension diversification in legal information retrieval. In *Proceedings of the 17th International Web Information Systems Engineering – WISE '16*, pages 174–189. Springer, 2016. `doi:10.1007/978-3-319-48740-3_12`.

5. Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. Diversifying the legal order. In Lazaros Iliadis and Ilias Maglogiannis, editors, *IFIP Advances in Information and Communication Technology*, pages 499–509. Springer Nature, 2016. `doi:10.1007/978-3-319-44944-9_44`.

6. Marios Koniaris, George Papastefanatos, and Yannis Vassiliou. Towards automatic structuring and semantic indexing of legal documents. In *Proceedings of the 20th Pan-Hellenic Conference on Informatics, PCI '16*. ACM Press, 2016. `doi:10.1145/3003733.3003801`.

7. Giorgos Giannopoulos, Marios Koniaris, Ingmar Weber, Alejandro Jaimes, and Timos Sellis. Algorithms and criteria for diversification of news article comments. *Journal of Intelligent Information Systems*, 44(1):1–47, 2015. `doi:10.1007/s10844-014-0328-1`.

8. Marios Koniaris, Giorgos Giannopoulos, Timos Sellis, and Yiannis Vasileiou. Diversifying Microblog Posts. In *Proceedings of the 15th International Web Information Systems Engineering - WISE '14*, pages 189–198. Springer International Publishing, 2014. `doi:10.1007/978-3-319-11746-1_14`.

9. Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. Legislation as a complex network: Modelling and analysis of European Union legal sources. In *Proceedings of the Twenty Seventh annual conference on Legal Knowledge and Information Systems - JURIX '14*, pages 143–152. IOS Press, 2014. `doi:10.3233/978-1-61499-468-8-143`.

Εργασιακή Εμπειρία

- 09/2015 - ..., Υπουργείο Παιδείας, Έρευνας και Θρησκευμάτων
  Γενική Γραμματεία Έρευνας και Τεχνολογίας (ΓΓΕΤ)
  Διεύθυνση Διεθνούς Επιστημονικής και Τεχνολογικής Συνεργασίας
  Τμήμα Ευρωπαϊκής Ένωσης

  - Παρακολούθηση των δραστηριοτήτων της Ευρωπαϊκής Ένωσης σε θέματα Έρευνας και Καινοτομίας - Ορίζοντας 2020
  - Συμμετοχή στη διαμόρφωση εθνικών θέσεων στο Συμβούλιο της Ευρωπαϊκής Ένωσης στον τομέα της Έρευνας και Καινοτομίας
  - Συμμετοχή σε ομάδες εμπειρογνωμόνων
  - Διαχείριση έργων κοινών προγραμμάτων, συγχρηματοδοτούμενων από Κοινοτικούς και Εθνικούς πόρους

- 10/2005 - 08/2015, Υπουργείο Παιδείας, Έρευνας και Θρησκευμάτων
  Εκπαιδευτικός Πληροφορικής Δευτεροβάθμιας Εκπαίδευσης

- 01/2004 – 10/2005 , EFG EUROBANK
  Διεύθυνση Υποστήριξης Κεντρικών Συστημάτων - Μηχανικός Πληροφορικής

  - Υπεύθυνος για την εύρυθμη λειτουργία των Κεντρικών Συστημάτων και του Δικτύου Καταστημάτων της Τράπεζας
  - Παραμετροποίηση και βελτιστοποίηση λειτουργίας των κεντρικών πληροφοριακών συστημάτων της Τράπεζας
  - Χρησιμοποιούμενες τεχνολογίες: z OS – OS 390, DB2 RDBMS, CICS Transaction Server

- 10/2005 – 1/2009, ΔΙΟΛΚΟΣ Α.Ε
  Τεχνικός Σύμβουλος (Εξωτερικός Συνεργάτης)

  - Ανάλυση και ανάπτυξη ολοκληρωμένων πληροφοριακών συστημάτων σε περιβάλλοντα J2EE

- 07/2003 – 09/2005, ΔΙΟΛΚΟΣ Α.Ε
  Technical Project Manager

  - Μελέτη, ανάλυση και ανάπτυξη ολοκληρωμένων πληροφοριακών συστημάτων εξειδικευόμενα στους τομείς των τραπεζο – ασφαλιστικών προϊόντων και την μέσω διαδικτύου διάθεση τους στα δίκτυα συνεργατών (B2B) και στον καταναλωτή (B2C)
  - Χρησιμοποιούμενες τεχνολογίες: WebServices, Java, Oracle 10g, Oracle JDeveloper, XML

– 11/1998 – 03/2001, Δημοσιογραφικός Οργανισμός Λαμπράκη ΔΟΛ Α.Ε
MULTIMEDIA A.E. - Μηχανικός Πληροφορικής

  – Ανάπτυξη ολοκληρωμένων πληροφοριακών συστημάτων αυτοματοποίησης και υποβοήθησης της παραγωγικής διαδικασίας του δημοσιογραφικού οργανισμού
  – Εκσυγχρονισμός εκδοτικών διαδικασιών για την ανάπτυξη νέων προϊόντων ηλεκτρονικών εκδόσεων
  – Ανάπτυξη εφαρμογών και συστημάτων με χρήση προηγμένων τεχνολογιών για βάσεις δεδομένων
  – Δημιουργία δυναμικών δικτυακών τόπων με χρήση τεχνολογιών αιχμής
  – Συντονισμός των τεχνικών θεμάτων μεταξύ των ποικίλλων πόρων του δημοσιογραφικού οργανισμού
  – Συμμετοχή στη διαδικασία καθορισμού πληροφοριακών απαιτήσεων και αποδοχής ποιότητας των διαφόρων συστημάτων
  – Χρησιμοποιούμενες τεχνολογίες: Java, C, Pro C, SQL, Oracle RDBMS, Oracle Developer, SQL Server, XML, HTML, CGI, JavaScript,Unix (AIX), Linux

– 11/1998 – 03/2001, Δημοσιογραφικός Οργανισμός Λαμπράκη ΔΟΛ Α.Ε
Αρχείο ΔΟΛ - Μηχανικός Ανάπτυξης Λογισμικού

  – Ερευνητικό έργο για την ηλεκτρονική μετατροπή και αρχειοθέτηση του αρχείου του δημοσιογραφικού οργανισμού
  – Χρησιμοποιούμενες τεχνολογίες: OCR (πολυτονικά κείμενα), Image and Hypertext Representation

Εκπαιδευτική εμπειρία

– 2010-2016: Επικουρικό εκπαιδευτικό έργο στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Ε.Μ.Π. για το μάθημα:

  – Ανάλυση και Σχεδιασμός Πληροφοριακών Συστημάτων (Καθ. Ιωάννης Βασιλείου)

– 03/2005 – 06/2005 & 10/2003 – 06/2004: Εργαστηριακός Συνεργάτης στο Τμήμα Πληροφορικής του ΤΕΙ Αθηνών στα κάτωθι γνωστικά αντικείμενα:

  – Γλώσσα Προγραμματισμού C (Καθ. Αλέξανδρος Τομαράς)
  – Γλώσσα Προγραμματισμού C++ (Καθ. Αλέξανδρος Τομαράς)

Ξένες Γλώσσες

– Αγγλικά: Proficiency (University of Cambridge)

– Γαλλικά - Ιταλικά: Μέτρια Κατανόηση