ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εμπλουτισμός Οντολογιών με Τεχνικές Μηχανικής Μάθησης.

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κυριακή Δ. Ζαφειρούδη

Επιβλέπων : Γεώργιος Στάμου

Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2016

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εμπλουτισμός Οντολογιών με Τεχνικές Μηχανικής Μάθησης.

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κυριακή Δ. Ζαφειρούδη

Επιβλέπων : Γεώργιος Στάμου

Επίκουρος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 27η Ιουνίου 2016.

.....	.....	.....
ΓΕΩΡΓΙΟΣ ΣΤΑΜΟΥ	ΣΤΕΦΑΝΟΣ ΚΟΛΛΙΑΣ	ΑΝΔΡΕΑΣ-ΓΕΩΡΓΙΟΣ ΣΤΑΦΥΛΟΠΑΤΗΣ
Επίκουρος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2016

.....

Κυριακή Δ. Ζαφειρούδη

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Καθώς ο Σημασιολογικός Ιστός κατέχει ύψιστη σημασία, δεδομένου ότι παρέχει μια κατανοητή από τις μηχανές προσέγγιση για χειρισμό πληροφοριών, οι Οντολογίες παίζουν το δικό τους ρόλο στην σύνθεση αυτού του νέου ιστού. Σε αντίθεση με τις βάσεις δεδομένων που παράγονται εδώ και πολλά χρόνια, υπάρχουν μόνο λίγες Οντολογίες για διαφόρους σκοπούς και οι μηχανικοί γνώσης δυσκολεύονται να συμβαδίσουν με το ρυθμό των σημερινών αναγκών. Ευτυχώς, στις μέρες μας υπάρχει τεράστια ανάπτυξη στις τεχνικές Μηχανικής Μάθησης (Machine Learning) που επιτρέπουν στις μηχανές να διαχειρίζονται πληροφορίες και να παράγουν συσχετίσεις για την ευκολότερη κατανόηση και επεξεργασία των δεδομένων. Στην παρούσα διπλωματική, ένα νέο μοντέλο προτείνεται για τον εμπλουτισμό υπαρχουσών Οντολογιών με την χρήση τεχνικών Μηχανικής Μάθησης. Πιο συγκεκριμένα, χρησιμοποιούνται τεχνικές ταξινόμησης (classification) και παραγωγής κανόνων συσχέτισης (association rules). Μια Οντολογία εμπλουτίζεται μέσω μοντέλων Μηχανικής Μάθησης τα οποία είχαν προηγουμένως εκπαιδευτεί με τη χρήση άλλων διαφορετικών Οντολογιών. Επιπλέον, παράγονται κανόνες συσχέτισης ανάλογα με τις έννοιες (concepts) και τις ιδιότητες (properties) που συνθέτουν τη δοσμένη Οντολογία, οι οποίοι εισάγονται στη συνέχεια στην αρχική Οντολογία με τη μορφή αξιωμάτων. Οι διαδικασίες της ταξινόμησης και της εκμάθησης κανόνων συσχέτισης διευκολύνεται από την χρήση του συστήματος Weka που παρέχει μια ποικιλία από υλοποιήσεις αλγορίθμων μηχανικής μάθησης. Για την επίτευξη του εμπλουτισμού της Οντολογίας γράφτηκε ένα πρόγραμμα σε Java και εκτελέστηκαν αρκετά υποσχόμενα πειράματα. Στις περισσότερες των περιπτώσεων ο εμπλουτισμός της Οντολογίας ήταν επιτυχής, καθώς νέοι ισχυρισμοί εννοιών και ρόλων προστέθηκαν στην Οντολογία μετά την εφαρμογή ταξινόμησης, βασισμένης σε άλλη δοσμένη οντολογία, καθώς και επιπρόσθετοι κανόνες συσχέτισης παράχθηκαν που προμήθευσαν την Οντολογία με νέα αξιώματα.

Λέξεις-κλειδιά: Οντολογίες, Μηχανική Μάθηση, Ταξινόμηση, Κανόνες Συσχέτισης, Weka.

Abstract

On the grounds that the Semantic Web is of utmost, since it provides a machine-understandable approach for information handling, Ontologies play their part in composing this new web. In contrast with databases that have been produced throughout many years, there are only a few Ontologies for each purpose, and Knowledge Engineers find it hard to keep up with the pace of today's needs. Luckily, nowadays there is also a huge development of Machine Learning techniques that enable machines to process information and produce associations to comprehend and edit data more easily. In the presented thesis, a new model is proposed to enrich existing Ontologies by the use of Machine Learning techniques. More accurately, classification and association rule techniques are utilized. An Ontology is enriched by Machine Learning models that were previously trained with the use of some other acquired Ontologies. Moreover, association rules are produced depending on the concepts and the properties of the given Ontology that are later added to the Ontology in forms of axioms. The procedures of classification and association rule learning are facilitated by the Weka system that implements a variety of Machine Learning algorithms. A Java program was written to carry out the enrichment of the Ontology and various promising experiments were conducted. In most cases the enrichment of the Ontology was successful, as new class and object property assertions were added to the Ontology after applying classification based on another given ontology and additional association rules were produced that provided new axioms to the Ontology.

Keywords: Ontologies, Machine Learning, Classification, Association Rules, Weka.

Περιεχόμενα

Κεφάλαιο 1. Εισαγωγή

Η έννοια του Σημασιολογικού Ιστού, όπως προτάθηκε από τον Tim Berners-Lee, εισήχθη σε μια προσπάθεια να καλυφθεί η ανάγκη της αναζήτησης και της ερμηνείας κατανοητών από τις μηχανές δεδομένων στον Ιστό. Στο Semantic Web Roadmap [1], ο Berners-Lee αναφέρει πως:

> "The Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help. One of the major obstacles to this has been the fact that most information on the Web is designed for human consumption, and even if it was derived from a database with well-defined meanings (in at least some terms) for its columns, that the structure of the data is not evident to a robot browsing the web. Leaving aside the artificial intelligence problem of training machines to behave like people, the Semantic Web approach instead develops languages for expressing information in a machine process-able form."

Σύμφωνα με το Semantic Web Activity Statement [2] οι κύριες τεχνολογίες του Σημασιολογικού Ιστού αποτελούνται από ένα σύνολο από πολυεπίπεδες προδιαγραφές και τα τρέχοντα στοιχεία είναι τα ακόλουθα:

- Resource Description Framework (RDF) Core Model
- RDF Schema language
- Web Ontology language (OWL)
- Simple Knowledge Organization System (SKOS)

Η έννοια της οντολογίας είναι ένα ζωτικό μέρος του Σημασιολογικού Ιστού. Σήμερα, οι εφαρμογές που βασίζονται στις οντολογίες και στα πρότυπα που παρέχονται από το World Wide Web Consortium (W3C) παρουσιάζουν «έξυπνη» συμπεριφορά χάρη στη συμβολική αναπαράσταση γνώσης και τις προδιαγραφές λογικών σχέσεων μεταξύ των αντικειμένων. Στη φιλοσοφία, οντολογία είναι η φιλοσοφική μελέτη της ύπαρξης και συγκρότησης του Όντος, της φύσης και της ουσίας των Όντων. Στην επιστήμη των υπολογιστών και την πληροφορική, οντολογία είναι η τυπική ονοματοδοσία και ο ορισμός των τύπων, των ιδιοτήτων και των αλληλεξαρτήσεων των οντοτήτων που υπάρχουν πραγματικά ή ουσιαστικά για ένα συγκεκριμένο πεδίο του λόγου. Πρόκειται λοιπόν για μια πρακτική εφαρμογή της φιλοσοφικής οντολογίας, με μια ταξινόμηση.

Ένας άλλος κοινός ορισμός της οντολογίας προέρχεται από τον Gruber [3], όπου μια οντολογία περιγράφεται ως «μια ρητή προδιαγραφή μιας επίνοιας» (“an explicit specification of a conceptualization”). Ο ορισμός από τον Gruber αργότερα αναπτύχθηκε περαιτέρω από τους Studer et al. [4], που αναφέρουν πως «η οντολογία είναι μια ρητή, τυπική προδιαγραφή μια επίνοιας» (”an ontology is a formal, explicit specification of a shared conceptualization”).

Πιο πρακτικά, οι οντολογίες είναι τα μέσα για τις μηχανές ώστε να διαβάζουν και να κατανοούν την ανθρώπινη επικοινωνία. Όταν οι άνθρωποι επικοινωνούν, χρησιμοποιούν λέξεις που έχουν κάποιο νόημα για αυτούς, ή ακόμα και περισσότερα από ένα νοήματα, αλλά η σημασιολογίας της φράσης μέσα στην οποία βρίσκονται αυτές οι λέξεις επιτρέπουν στους ανθρώπους να καταλαβαίνουν ο ένας τον άλλον. Για τις μηχανές αυτή η διαδικασία δεν είναι τόσο απλή. Οι λέξεις είναι σειρές συνεχόμενων χαρακτήρων που δεν αντιπροσωπεύουν κάποιο νόημα. Ως εκ τούτου, στην Τεχνητή Νοημοσύνη υπάρχει η ανάγκη για τις οντολογίες ώστε να τοποθετήσουν τις λέξεις σε μια δομή από έννοιες, όπου οι συνδέσεις μεταξύ των εννοιών που αντιστοιχούν σε λέξεις να περιγράφονται με έναν τυπικό και σαφή τρόπο.

Οι οντολογίες αποτελούνται από κλάσεις (classes ή sets), χαρακτηριστικά (attributes ή properties) και σχέσεις (relationships ή relations among class members). Μια οντολογία μαζί με ένα σύνολο από μεμονωμένα στιγμιότυπα των κλάσεων αποτελούν μια βάση γνώσης. Στην πραγματικότητα, είναι λεπτή η γραμμή που διαχωρίζει το που τελειώνει η οντολογία και ξεκινάει η βάση γνώσης.

Υπάρχει μια συνεχώς αυξανόμενη ανάγκη για οντολογίες σε διάφορα πεδία. Σε αντίθεση με τις βάσεις δεδομένων, δεν υπάρχει αφθονία οντολογιών. Ως εκ τούτου, η αναγκαιότητα για τον εμπλουτισμό των υφιστάμενων οντολογιών αναδύεται, καθώς η διαδικασία της δημιουργίας οντολογιών από το μηδέν είναι χρονοβόρα και επομένως δαπανηρή δραστηριότητα. Μια αυτοματοποιημένη ή έστω ήμι-αυτοματοποιημένη προσέγγιση στον εμπλουτισμό οντολογιών μπορεί να χρησιμοποιηθεί από τους μηχανικούς γνώσης στην κατασκευή νέων οντολογιών βασισμένων σε υπάρχουσες οντολογίες ή την επέκταση υφισταμένων. Αυτή η προσέγγιση διευκολύνεται με τη χρήση τεχνικών μηχανικής μάθησης και ειδικότερα τεχνικών εξόρυξης κειμένου και ταξινόμησης.

Η μηχανική μάθηση είναι μια κεντρική υποπεριοχή της τεχνητής νοημοσύνης που επιτρέπει στις μηχανές να εκπαιδεύονται από μόνες τους με καθόλου ή μερική επίβλεψη από τους ανθρώπους. Οι υπολογιστές είναι σε θέση να μαθαίνουν μόνοι τους από δεδομένα χωρίς να έχουν προγραμματιστεί ρητά και με λεπτομέρεια. Αυτή η ικανότητα προσαρμογής σε νέα δεδομένα είναι υψίστης σημασίας όταν πρόκειται για μεγάλο όγκο δεδομένων. Θα εκμεταλλευτούμε την βοήθεια της μηχανικής μάθησης στην αναγνώριση προτύπων στα δεδομένα και στην παρατήρηση ομοιοτήτων και συσχετίσεων αόρατων προς το ανθρώπινο μάτι.

Ο στόχος της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη ενός ευφυούς συστήματος ικανού να συνδυάζει δεδομένα ώστε να παράγει αυτόματα ένα πλουσιότερο σύνολο δεδομένων από αυτό της αρχική οντολογίας με τη χρήση τεχνικών μηχανικής μάθησης. Οι πληροφορίες που εκπροσωπούνται στο σώμα ισχυρισμών μιας οντολογίας είναι πιθανό να περιέχουν κρυφές σχέσεις οι οποίες θα μπορούσαν να προτείνουν νέους ισχυρισμούς ή υπαγωγές γενικών εννοιών, οι οποίες θα δημιουργούσαν μια πλουσιότερη βάση γνώσης. Τα οφέλη της μηχανικής μάθησης στην αυτόματη ταξινόμηση και παραγωγή κανόνων συσχέτισης εξερευνήθηκαν κατά τη διαδικασία ανακάλυψης κρυμμένων μοτίβων στα δεδομένα, με την εξαγωγή πληροφοριών που παρουσιάζονται στην οντολογία και την εφαρμογή τεχνικών μηχανικής μάθησης σε αυτά. Το αποτέλεσμα ήταν μια διαδικασία μηχανικής μάθησης που επιτρέπει τον εμπλουτισμό οντολογιών με τη χρήση φίλτρων, ταξινόμησης και κανόνων συσχέτισης.

Οι οντολογίες – μια προς εμπλουτισμό και μια ή περισσότερες που θα λειτουργήσουν ως μια ολοκληρωμένη και σαφώς ορισμένη βάση γνώσης που θα χρησιμοποιηθεί αργότερα για την εκπαίδευση του ταξινομητή (classifier) – που δέχεται το προτεινόμενο σύστημα ως είσοδο είναι σε μορφή OWL. Για την πραγματοποίηση του μοντέλου γράφτηκε ένα πρόγραμμα σε Java, το οποίο μετέτρεπε τα αρχεία OWL σε αρχεία ARFF όπου τα γνωρίσματα βρίσκονταν σε αντιστοιχία με τα στιγμιότυπα (individuals), τις κλάσεις (classes), τις ιδιότητες αντικειμένου (object properties) και τις ιδιότητες τύπου δεδομένων (data properties) των αρχικών οντολογιών, δημιουργώντας ένα σύνολο δεδομένων εκπαίδευσης (training dataset) και ένα δοκιμής (testing dataset) σε αντιστοιχία με τις δύο οντολογίες που αναφέρονται προηγουμένως. Μια διαδικασία προεπεξεργασίας (preprocessing) αναπτύχθηκε

ώστε να καταστούν τα δύο σύνολα δεδομένων συμβατά και να διευκολυνθεί η διεργασία ταξινόμησης. Οι πιο ταιριαστοί ταξινομητές για την επίτευξη του επιθυμητού σκοπού επιλέχθηκαν και εκπαιδεύτηκαν με τα δεδομένα. Στη συνέχεια, το σύνολο δεδομένων δοκιμής ταξινομήθηκε υπακούοντας στους ταξινομητές που παράχθηκαν. Επιπλέον, το σύνολο δεδομένων δοκιμής επέτρεψε την εξόρυξη κανόνων συσχέτισης με τον επακόλουθο εμπλουτισμό της αρχικής οντολογίας με ισχυρισμούς. Τέλος, το πρόγραμμα παράγει αξιώματα OWL από το ARFF αρχείο του προκύπτοντος συνόλου δεδομένων τα οποία προστίθενται στη συνέχεια στην αρχική οντολογία.

Το αρχικό τμήμα του Κεφαλαίου 2 ξεκινάει με μια εισαγωγή στις Περιγραφικές Λογικές και τις Οντολογίες μαζί με τους κανόνες και τα αξιώματα που συνθέτουν μια οντολογία, ακολουθούμενα από μία ματιά στη σημασιολογία πίσω από αυτά και την εξαγωγή συμπερασμάτων (reasoning). Στη συνέχεια, η επόμενη ενότητα θα ασχοληθεί με το θεωρητικό υπόβαθρο που απαιτείται για τη χρήση μηχανικής μάθησης. Τα διαφορετικά γνωρίσματα που δημιουργήθηκαν, το σύστημα Weka που χρησιμοποιήθηκε και οι διάφορες εργασίες μηχανικής μάθησης που εκτελέστηκαν θα παρουσιαστούν αναλυτικά.

Μια εξέταση σχετικών συστημάτων που έχουν αναπτυχθεί θα γίνει στο Κεφάλαιο 3, περιγράφοντας διάφορες διαδικασίες εξόρυξης πληροφοριών από οντολογίες, καθώς και για τον σκοπό της δημιουργίας οντολογιών, και προγράμματα που έχουν αναπτυχθεί πρωτύτερα και προτείνουν παρόμοια αποτελέσματα με την παρούσα διπλωματική.

Το προτεινόμενο μοντέλο που έχει αναπτυχθεί θα ακολουθήσει στο Κεφάλαιο 4, μαζί με την επιχειρηματολογία για τις αποφάσεις που λήφθηκαν, αναλυτική περιγραφή της προεπεξεργασίας των οντολογιών, λεπτομέρειες σχετικά με τις τεχνικές μηχανικής μάθησης που χρησιμοποιήθηκαν, καθώς και μια διεξοδική επίδειξη του εφαρμοσμένου συστήματος.

Τα πειραματικά αποτελέσματα θα παρουσιαστούν στο Κεφάλαιο 5, με μια συζήτηση σχετικά με το κατά πόσο τα αποτελέσματα είναι χρήσιμα για την προσθήκη καινούργιας γνώσης στην οντολογία, μια επισκόπηση του σχεδιασμού των οντολογιών που χρησιμοποιήθηκαν για την εκτέλεση του προγράμματος και τις ιδιαιτερότητες των προκυπτόντων οντολογιών.

Στο Κεφάλαιο 6 θα εξαχθούν συμπεράσματα από τα πειραματικά αποτελέσματα και θα αξιολογηθεί το έργο που επιτεύχθηκε ως προς το πώς και εάν θα μπορούσε να γενικευτεί ανεξάρτητα από την οντολογία, ολοκληρώνοντας με πιθανές μελλοντικές επεκτάσεις του μοντέλου.

Κεφάλαιο 2. Προκαταρκτικά

2.1. Οντολογίες

Σε αυτή την ενότητα παρέχεται μια εισαγωγή στις Περιγραφικές Λογικές, τον τρόπο με τον οποίο μοντελοποιείται η γνώση στις Περιγραφικές Λογικές καθώς και τα πιο σημαντικά χαρακτηριστικά μοντελοποίησης τους, με σκοπό την κατάληξη στην ομοιότητά και τη χρήση τους στις Οντολογίες.

2.1.1. Περιγραφικές Λογικές

Οι *Περιγραφικές Λογικές* είναι μια οικογένεια γλωσσών αναπαράστασης γνώσης που χρησιμοποιούνται ευρέως στην μοντελοποίηση οντολογιών. Ο κύριος λόγος για την τόσο διαδεδομένη χρήση τους είναι πως παρέχουν τα θεμέλια για την OWL Web Ontology Language όπως έχει τυποποιηθεί από το World Wide Web Consortium

(W3C), παρότι έχουν χρησιμοποιηθεί στην αναπαράσταση γνώσης πολύ πριν την μοντελοποίηση οντολογιών στα πλαίσια του Σημασιολογικού Ιστού.

Το πιο σημαντικό χαρακτηριστικό που διαχωρίζει τις Περιγραφικές Λογικές από άλλες γλώσσες μοντελοποίησης, είναι πως οι Περιγραφικές Λογικές είναι λογικές, με την έννοια πως είναι εξοπλισμένες με τυπική σημασιολογία. Αυτή η σημασιολογία επιτρέπει στους ανθρώπους και τα πληροφοριακά συστήματα να ανταλλάσσουν οντολογίες σε Περιγραφικές Λογικές χωρίς ασάφειες ως προς το νόημά τους και να χρησιμοποιούν λογική επαγωγή για να καταλήξουν σε λογικές υποθέσεις από τα γεγονότα που αναφέρονται ρητά στην οντολογία. Αυτά τα συμπεράσματα προκύπτουν από την διαδικασία υπολογισμού που ονομάζεται συλλογιστική (reasoning).

Οι Περιγραφικές Λογικές (ΠΛ) παρέχουν τα μέσα για την μοντελοποίηση σχέσεων μεταξύ οντοτήτων σε μια περιοχή ενδιαφέροντος. Στις Περιγραφικές Λογικές, τα στοιχεία μιας περιοχής ενδιαφέροντος διαρθρώνονται σε έννοιες (μοναδιαία κατηγορήματα – unary predicates), οι ιδιότητές τους καθορίζονται μέσω ρόλων (δυαδικά κατηγορήματα – binary predicates) και τα ξεχωριστά άτομα αναπαρίστανται από ονόματα ατόμων (σταθερές – constants). Σε αντίθεση με μια βάση δεδομένων η οποία περιγράφει πλήρως μια συγκεκριμένη κατάσταση, μια ΠΛ οντολογία αποτελείται από ένα σύνολο καταστάσεων, τα λεγόμενα αξιώματα, καθένα από τα οποία πρέπει να ισχύουν στην σκιαγραφημένη κατάσταση. Σύνθετες εκφράσεις εννοιών και ρόλων κατασκευάζονται (έννοιες και ρόλοι για απλότητα), ξεκινώντας από ένα σύνολο από ονόματα εννοιών και ρόλων, με την εφαρμογή κατάλληλων κατασκευαστών, όπου το σύνολο των διαθέσιμων κατασκευαστών εξαρτάται από τη συγκεκριμένη ΠΛ. Οι έννοιες και οι ρόλοι μπορούν να χρησιμοποιηθούν σε μια βάση γνώσης για την εξαγωγή γνώσης, τόσο σε επίπεδο σώματος ορολογίας (TBox – Terminological Box), όσο και σε επίπεδο σώματος ισχυρισμών (ABox – Assertional Box). Το TBox συνήθως αποτελείται από ένα σύνολο αξιωμάτων που δηλώνουν υπαγωγή ανάμεσα σε έννοιες και ρόλους. Στο ABox, τα αξιώματα αναφέρονται σε γνώση σχετική με επονομαζόμενα άτομα. Οι Περιγραφικές Λογικές υποστηρίζονται από μηχανισμούς εξαγωγής συμπερασμάτων (reasoning), όπως είναι ο έλεγχος ικανοποιησιμότητας και η απάντηση ερωτημάτων, τα οποία στηρίζονται στην βασισμένη στην λογική σημασιολογία τους.

Αξιώματα ABox

Τα ABox αξιώματα αποτελούνται από ισχυρισμούς εννοιών (concept assertions) και ρόλων (role assertions). Οι *ισχυρισμοί εννοιών* περιγράφουν τη συμμετοχή αντικειμένων σε έννοιες, όπως

$$Koala(Baboo)$$

το οποίο κάνει τον ισχυρισμό πως ο Baboo είναι κοάλα ή, πιο συγκεκριμένα, πως το άτομο με την ονομασία $Baboo$ είναι στιγμιότυπο της έννοιας $Koala$.

Οι *ισχυρισμοί ρόλων* περιγράφουν πως ένα ζευγάρι αντικειμένων συνδέεται μέσω ενός ρόλου, όπως

$$hasChild(Baboo, Bambi)$$

το οποίο κάνει τον ισχυρισμό πως ο Bambi είναι παιδί του Bamboo ή, πιο συγκεκριμένα, πως το άτομο με την ονομασία $Baboo$ συμμετέχει στη σχέση που εκπροσωπείται από το $hasChild$ με το άτομο που ονομάζεται $Bambi$.

Στην ανθρώπινη διαίσθηση, είναι προφανές πως ο $Baboo$ και ο $Bambi$ δεν είναι το ίδιο άτομο, αλλά οι Περιγραφικές Λογικές δεν κάνουν την υπόθεση του μοναδικού ονόματος (unique name assumption), οπότε διαφορετικά ονόματα ατόμων μπορεί να αναφέρονται στην ίδια οντότητα εκτός και αν αναφέρεται ρητά το αντίθετο. Για τον ανώτερο λόγο, μπορούμε να ορίσουμε πως δύο άτομα είναι διαφορετικά με τον ισχυρισμό της ανισότητας των ατόμων, όπως

$$Baboo \not\approx Bambi$$

το οποίο κάνει τον ισχυρισμό πως ο $Baboo$ και ο $Bambi$ είναι όντως διαφορετικά άτομα. Από την άλλη, ο ισχυρισμός της ισότητας των ατόμων, όπως

$$Baboo \approx Bambi$$

κάνει τον ισχυρισμό πως δύο διαφορετικά ονόματα ατόμων αναφέρονται στην ίδια οντότητα.

Αξιώματα TBox

Τα TBox αξιώματα αποτελούνται από υπαγωγές εννοιών, ισοδυναμία εννοιών και ιδιότητες ρόλων. Οι *υπαγωγές εννοιών* περιγράφουν πως μια έννοια υπάγεται σε μία άλλη έννοια, όπως

$$Koala \sqsubseteq Animal$$

το οποίο κάνει τον ισχυρισμό πως κάθε κοάλα είναι επίσης και ζώο ή, πιο συγκεκριμένα, πως η έννοια $Koala$ υπάγεται στην έννοια $Animal$.

Η *ισοδυναμία εννοιών* περιγράφει πως δύο έννοιες έχουν τα ίδια στιγμιότυπα, όπως

$$Person \equiv Human$$

το οποίο κάνει τον ισχυρισμό πως κάθε στιγμιότυπο της έννοιας $Person$ είναι επίσης στιγμιότυπο της έννοιας $Human$.

Ξένες έννοιες μπορούν να δηλωθούν όταν δύο έννοιες δεν μπορούν να ισχύουν ταυτόχρονα για τα ίδια στιγμιότυπα, όπως

$$Disjoint(Koala, Person)$$

το οποίο κάνει τον ισχυρισμό πως η έννοια $Koala$ και η έννοια $Person$ είναι ξένες μεταξύ τους.

Οι *υπαγωγές ρόλων* περιγράφουν πως ένας ρόλος είναι υπο-ρόλος ενός άλλου, όπως

$$hasChild \sqsubseteq hasOffspring$$

το οποίο κάνει τον ισχυρισμό πως κάθε στιγμιότυπο που έχει ένα παιδί, έχει επιπλέον και έναν απόγονο, ή, πιο συγκεκριμένα, πως η σχέση που αναπαρίσταται από το $hasChild$ είναι ένας υπο-ρόλος της σχέσης που αναπαρίσταται από το $hasOffspring$.

Η *σύνθεση ρόλων* μπορεί να χρησιμοποιηθεί σε αξιώματα υπαγωγής ρόλων για την περιγραφή ρόλων σε αξιώματα υπαγωγής ρόλων, όπως

$$hasChild \circ hasChild \sqsubseteq hasGrandchild$$

το οποίο κάνει τον ισχυρισμό πως ένα άτομο που έχει κάποιο άλλο άτομο ως παιδί, το οποίο έχει επίσης παιδί, έχει συμπερασματικά και εγγόνι. Σημειώνεται πως η σύνθεση ρόλων μπορεί να εμφανιστεί μόνο στην αριστερή πλευρά σύνθετων αξιωμάτων υπαγωγής ρόλων.

Ξένοι ρόλοι μπορούν να δηλωθούν όταν δύο ρόλοι δεν μπορούν να ισχύουν ταυτόχρονα για τα ίδια άτομα, όπως

$$Disjoint(hasChild, hasGrandchild)$$

το οποίο κάνει τον ισχυρισμό πως το $hasChild$ και το $hasGrandchild$ είναι ξένοι ρόλοι.

Οι κατασκευαστές εννοιών και ρόλων μας επιτρέπουν να διαμορφώσουμε σύνθετες έννοιες και ρόλους.

Δυαδικοί Κατασκευαστές Εννοιών

Βασικές δυαδικές εργασίες επιτυγχάνονται με τη χρήση των δυαδικών κατασκευαστών εννοιών (Boolean concept constructors). Η *τομή* (intersection or conjunction) περιγράφει ένα σύνολο ατόμων που αποτελείται από τα ακριβή άτομα που είναι στιγμιότυπα όλων των εννοιών που συμμετέχουν στην τομή, όπως

$$Animal \sqcap FourFooted$$

το οποίο αναφέρεται σε όλα τα στιγμιότυπα τα οποία είναι ταυτόχρονα ζώα και τετράποδα. Σύνθετες έννοιες μπορούν να χρησιμοποιηθούν σε αξιώματα σαν ατομικές έννοιες, όπως $Quadruped \equiv Animal \sqcap FourFooted$.

Η *ένωση* (union or disjunction) περιγράφει ένα σύνολο από άτομα που αποτελείται από όλα τα άτομα που είναι στιγμιότυπα τουλάχιστον σε μία από τις έννοιες που συμμετέχουν στην ένωση – το δυαδικό ανάλογο της τομής, όπως

$$FemaleKoala \sqcup MaleKoala$$

το οποίο περιγράφει όλα τα στιγμιότυπα τα οποία είναι είτε θηλυκά είτε αρσενικά κοάλα, όπως $Koala \equiv FemaleKoala \sqcup MaleKoala$.

Η *άρνηση* (complement or negation) περιγράφει ένα σύνολο ατόμων τα οποία δεν είναι στιγμιότυπα μιας συγκεκριμένης έννοιας, όπως

$$\neg Female$$

το οποίο περιγράφει όλα τα στιγμιότυπα που δεν είναι θηλυκά (στιγμιότυπα που είναι αρσενικά).

Η *καθολική έννοια* (top concept) $\top$ είναι μια ειδική έννοια της οποίας στιγμιότυπα είναι κάθε άτομο, όπως

$$\top \equiv C \sqcup \neg C$$

για μια αυθαίρετη έννοια C.

Μια άλλη ειδική έννοια είναι η *κενή έννοια* $\bot$ η οποία δεν έχει κανένα άτομο ως στιγμιότυπο, το δυαδικό ανάλογο του $\top$, όπως

$$\bot \equiv C \sqcap \neg C$$

για μια αυθαίρετη έννοια C.

Περιορισμοί Ρόλων

Οι Περιγραφικές Λογικές έχουν την ικανότητα να διαμορφώνουν δηλώσεις που συνδέουν έννοιες με ρόλους. Ο *υπαρξιακός περιορισμός* (existential restriction) είναι μια σύνθετη έννοια που περιγράφει ένα σύνολο ατόμων που είναι συνδεδεμένοι μέσω ενός συγκεκριμένου ρόλου με κάποιο άλλο άτομο, όπως

$$Parent \equiv \exists hasChild.\top$$

το οποίο κάνει τον ισχυρισμό πως το σύνολο των ατόμων που είναι γονείς βρίσκονται στη σχέση που αναπαρίσταται από το $hasChild$ με ένα άλλο άτομο το λιγότερο (στιγμιότυπο του $\top$).

Ο *περιορισμός τιμής* (universal restriction) περιγράφει ένα σύνολο ατόμων που είναι συνδεδεμένοι μέσω ενός συγκεκριμένου ρόλου με ένα συγκεκριμένο άτομο, όπως

$$\forall hasChild.Female$$

το οποίο περιγράφει ένα σύνολο από άτομα που βρίσκονται στη σχέση που αναπαρίσταται από το $hasChild$ με ένα άλλο άτομο το οποίο είναι στιγμιότυπο του $Female$. Να σημειωθεί πως αυτό το σύνολο περιέχει επίσης και τα άτομα που δεν έχουν καθόλου παιδιά.

Ο υπαρξιακός περιορισμός και ο περιορισμός τιμής σε συνδυασμό με την καθολική έννοια χρησιμοποιούνται για να εκφράσουν πεδίο ορισμού και τιμών στους ρόλους, όπως

$$\exists hasChild.\top \sqsubseteq \text{Parent}$$

το οποίο κάνει τον ισχυρισμό πως το πεδίο ορισμού του $hasChild$ περιορίζεται σε στιγμιότυπα της έννοιας $Parent$, και επιπλέον όπως

$$\top \sqsubseteq \forall \text{hasChild}.\,(\text{Female} \sqcup \text{Male})$$

το οποίο κάνει τον ισχυρισμό πως το πεδίο τιμών του $hasChild$ περιορίζεται σε στιγμιότυπα είτε της έννοιας $Female$ είτε της έννοιας $Male$.

Ο *περιορισμός πληθικότητας* (number restriction) επιτρέπει την οριοθέτηση του αριθμού των ατόμων που μπορούν να είναι προσβάσιμα μέσω κάποιου δοσμένου ρόλου, όπως ο το-λιγότερο (at-least) περιορισμός

$$\geq 3\ hasChild.Female$$

ο οποίος περιγράφει ένα σύνολο από άτομα που έχουν το λιγότερο τρείς κόρες. Υπάρχει επίσης ο το-πολύ (at-most) περιορισμός, όπως

$$\leq 3\ hasChild.Male$$

ο οποίος περιγράφει ένα σύνολο από άτομα τα οποία έχουν το πολύ τρείς γιούς.

Κατασκευαστές Ρόλων

Σύνθετοι ρόλοι μπορούν να διαμορφωθούν με μερικούς άλλους κατασκευαστές. Οι *αντίστροφοι ρόλοι* χρησιμοποιούνται για να υποδείξουν πως ένας ρόλος είναι το δυαδικό ανάλογο ενός άλλου, όπως

$$hasChild \equiv hasParent^{-}$$

το οποίο κάνει τον ισχυρισμό πως αν ένα άτομο βρίσκεται σε σχέση που αναπαρίσταται από το $hasChild$ με κάποιο άλλο άτομο, τότε το άλλο άτομο βρίσκεται σε σχέση που αναπαρίσταται από το $hasParent^{-}$ με το πρώτο άτομο, όπου ο ρόλος $hasParent^{-}$ αναπαριστά το αντίθετο του $hasParent$.

DL-Lite

Όμως για ποιες γλώσσες οντολογιών μπορούμε να απαντήσουμε εκφραστικά ερωτήματα (queries) πάνω σε μία οντολογία αποτελεσματικά; Η DL-Lite [5] είναι μία οικογένεια Περιγραφικών Λογικών βελτιστοποιημένη ως προς την ισορροπία ανάμεσα στην εκφραστικότητα και την πολυπλοκότητα των δεδομένων και είναι ειδικά προσαρμοσμένη να συλλάβει βασικές γλώσσες οντολογιών. Όπως συνηθίζεται στις Περιγραφικές Λογικές, η DL-Lite επιτρέπει την αναπαράσταση μιας περιοχής ενδιαφέροντος από πλευράς εννοιών, δηλώνοντας σύνολα αντικειμένων, και από πλευράς ρόλων, δηλώνοντας δυαδικές σχέσεις μεταξύ αντικειμένων.

Για τους σκοπούς της παρούσας διπλωματικής, δεν θεωρήθηκε σκόπιμο να περιγραφούν οι ιδιαιτερότητες κάθε επιμέρους τμήματος της οικογένειας DL-Lite. Ακολούθως παρουσιάζεται η σύνταξη και η σημασιολογία της συλλογιστικής στην DL-Lite.

Υποθέτουμε πως η DL-Lite περιέχει ονόματα ατόμων $a_0, a_1, \dots$, ονόματα εννοιών $A_0, A_1, \dots$, και ονόματα ρόλων $P_0, P_1, \dots$. Σύνθετοι ρόλοι R και έννοιες C αυτής της γλώσσας ορίζονται ως ακολούθως:

$$R ::= P_k | P_k^-$$

$$B ::= \bot \ | A_k | \geq qR$$

$$C ::= B | \neg C | C_1 \sqcap C_2$$

όπου το q είναι ένας θετικός ακέραιος. Οι έννοιες της μορφής B θα ονομάζονται βασικές.

Ένα DL-Lite TBox, $\mathcal{T}$, είναι ένα πεπερασμένο σύνολο από αξιώματα υπαγωγής εννοιών και ρόλων (υπαγωγή εννοιών και ρόλων για απλότητα – concept and role inclusions) της μορφής:

$$C_1 \sqsubseteq C_2$$

$$R_1 \sqsubseteq R_2$$

και ένα ABox, $\mathcal{A}$, είναι ένα πεπερασμένο σύνολο ισχυρισμών της μορφής:

$$A_k(a_i)$$

$$\neg A_k(a_i)$$

$$P_k(a_i, a_j)$$

$$\neg P_k(a_i, a_j)$$

Μαζί, το TBox $\mathcal{T}$ και το ABox $\mathcal{A}$ απαρτίζουν την DL-Lite βάση γνώσης (knowledge base) $\mathcal{K} = (\mathcal{T}, \mathcal{A})$. Στα ακόλουθα, δηλώνεται με το $role(\mathcal{K})$ το σύνολο των ονομάτων των ρόλων που προκύπτουν στα $\mathcal{T}$ και $\mathcal{A}$, με $role^{\pm}(\mathcal{K})$ το σύνολο $\{P_k, P_k^- | P_k \in role(\mathcal{K})\}$, και με $ob(\mathcal{A})$ το σύνολο των ονομάτων των ατόμων στο $\mathcal{A}$. Για ένα ρόλο R, ορίζεται:

$$inv(R) = \begin{cases} P_k^-, \alpha\nu\ R = P_k \\ P_k, \alpha\nu\ R = P_k^- \end{cases}$$

Όπως συνηθίζεται στις Περιγραφικές Λογικές, μια ερμηνεία, $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, αποτελείται από ένα μη κενό πεδίο $\Delta^{\mathcal{I}}$ και μια συνάρτηση ερμηνείας $\cdot^{\mathcal{I}}$ που αντιστοιχίζει κάθε όνομα ατόμου a_i σε ένα στοιχείο $a_i^{\mathcal{I}} \in \Delta^{\mathcal{I}}$, κάθε όνομα έννοιας A_k σε ένα υποσύνολο $A_k^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ του πεδίου, και κάθε όνομα ρόλου P_k σε μία δυαδική σχέση $P_k^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ πάνω στην περιοχή. Εκτός και αν αναφέρεται διαφορετικά, υιοθετούμε την υπόθεση μοναδικού ονόματος (unique name assumption):

$$a_i^{\mathcal{I}} \neq a_j^{\mathcal{I}} \; για \; κάθε \; i \neq j$$

Ωστόσο, πρέπει πάντοτε να αναφέρεται ποια από τα αποτελέσματα εξαρτώνται από την υπόθεση μοναδικού ονόματος και ποια όχι, και όταν εξαρτώνται από αυτή την υπόθεση, να αναφέρονται οι συνέπειες της εγκατάλειψής της.

Οι κατασκευαστές ρόλων και εννοιών ερμηνεύονται στην $\mathcal{I}$ με τον τυποποιημένο τρόπο:

$(P_k^-)^{\mathcal{I}}$	$= \{(y,x) \in \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}} \mid (x,y) \in P_k^{\mathcal{I}}\},$	(αντίστροφος ρόλος)
$\bot^{\mathcal{I}}$	$= \emptyset,$	(κενό σύνολο)
$(\geq q \; R)^{\mathcal{I}}$	$= \{x \in \Delta^{\mathcal{I}} \mid \#\{y \in \Delta^{\mathcal{I}} \mid (x,y) \in R^{\mathcal{I}}\} \geq q\},$	(το λιγότερο q Rδιάδοχοι)
$(\neg C)^{\mathcal{I}}$	$= \Delta^{\mathcal{I}} \backslash C^{\mathcal{I}},$	(δεν ανήκει στο C)
$(C_1 \sqcap C_2)^{\mathcal{I}}$	$= C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}},$	(ανήκει στο C_1 και στο C_2)

όπου το $\#X$ δηλώνει την πληθυκότητα του X. Θα γίνεται χρήση καθιερωμένων συντομεύσεων όπως

$$C_1 \sqcup C_2 = \neg(\neg C_1 \sqcap \neg C_2)$$

$$\top = \neg\bot$$

$$\exists R = (\geq 1R)$$

$$\leq q \; R = \neg(\geq q + 1 \; R)$$

Οι έννοιες της μορφής $\leq q \; R$ και $\geq q \; R$ ονομάζονται περιορισμοί πληθυκότητας, και αυτές της μορφής $\exists R$ ονομάζονται υπαρξιακοί περιορισμοί.

Η σχέση ικανοποίησης $\vDash$ είναι εξίσου καθιερωμένη:

$$\mathcal{I} \vDash C_1 \sqsubseteq C_2 \; \alpha\nu\nu \; C_1^{\mathcal{I}} \subseteq C_2^{\mathcal{I}}$$

$$\mathcal{I} \vDash R_1 \sqsubseteq R_2 \; \alpha\nu\nu \; R_1^{\mathcal{I}} \subseteq R_2^{\mathcal{I}}$$

$$\mathcal{I} \vDash A_k(a_i) \; \alpha\nu\nu \; a_i^{\mathcal{I}} \in A_k^{\mathcal{I}}$$

$$\mathcal{I} \vDash P_k(a_i, a_j) \; \alpha\nu\nu \; (a_i^{\mathcal{I}}, a_j^{\mathcal{I}}) \in P_k^{\mathcal{I}}$$

$$\mathcal{I} \vDash \neg A_k(a_i) \; \alpha\nu\nu \; a_i^{\mathcal{I}} \notin A_k^{\mathcal{I}}$$

$$\mathcal{I} \vDash \neg P_k(a_i, a_j) \; \alpha\nu\nu \; (a_i^{\mathcal{I}}, a_j^{\mathcal{I}}) \notin P_k^{\mathcal{I}}$$

Μια βάση γνώσης $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ λέγεται ικανοποιήσιμη (ή συνεπής) αν υπάρχει μια ερμηνεία $\mathcal{I}$, που να ικανοποιεί όλα τα μέλη του $\mathcal{T}$ και του $\mathcal{A}$. Σε αυτή την περίπτωση γράφουμε $\mathcal{I} \vDash \mathcal{K}$ (καθώς επίσης $\mathcal{I} \vDash \mathcal{T}$ και $\mathcal{I} \vDash \mathcal{A}$) και λέμε πως η $\mathcal{I}$ είναι ένα μοντέλο της $\mathcal{K}$ (και των $\mathcal{T}$ και $\mathcal{A}$).

2.1.2. Web Ontology Language

Η τρέχουσα έκδοση της Web Ontology Language είναι η OWL 2, μια σύσταση του W3C από τον Οκτώβριο 2009 [6]. Η OWL 2 βασίζεται σε γνωστές έννοιες και συμπεράσματα που αναφέρθηκαν προηγουμένως στις Περιγραφικές Λογικές. Όπως και οι Περιγραφικές Λογικές, η OWL 2 είναι μια γλώσσα αναπαράστασης γνώσης σχετικής με άτομα, ομάδες ατόμων και σχέσεις μεταξύ ατόμων. Για την αναπαράσταση γνώσης στην OWL 2, οι ακόλουθες έννοιες είναι θεμελιώδεις:

- Αξιώματα - Axioms: δηλώσεις που μια OWL οντολογία εκφράζει και ισχυρίζεται πως ισχύουν – αυτές περιλαμβάνουν τη συνολική θεωρία που περιγράφει η οντολογία στην περιοχή εφαρμογής της
- Οντότητες - Entities: έννοιες που αναφέρονται σε αντικείμενα του πραγματικού κόσμου – άτομα, κλάσεις, ιδιότητες.
- Εκφράσεις - Expressions: συνδυασμός των οντοτήτων για τον σχηματισμό σύνθετων εννοιών που επιτυγχάνεται συνδυάζοντας βασικές οντότητες με τη χρήση κατασκευαστών.

Τα αντικείμενα συμβολίζονται ως άτομα, οι κατηγορίες ως κλάσεις και οι σχέσεις ως ιδιότητες. Στην OWL 2 δεν υπάρχει μόνο ένας τύπος ιδιοτήτων: οι ιδιότητες αντικειμένων (object properties) δημιουργούν μια σχέση μεταξύ δύο ατόμων, ενώ οι ιδιότητες τύπων δεδομένων (datatype/data properties) συνδέουν ένα άτομο με μία τιμή τύπου δεδομένων. Για την κωδικοποίηση πληροφοριών σχετικές με την ίδια την οντολογία χρησιμοποιούνται οι ιδιότητες σχολιασμού (annotation properties), οι οποίες δεν έχουν καμία επίδραση στις πτυχές συλλογιστικής μίας οντολογίας.

Για τους σκοπούς της παρούσας διπλωματικής, θα χρησιμοποιηθεί η Fynctional-Style Syntax σύνταξη, η οποία έχει σχεδιαστεί για να είναι ευκολότερη για σκοπούς προσδιορισμών και για να παρέχει μια βάση για την εφαρμογή εργαλείων της OWL 2, όπως APIs και reasoners. Ακολούθως, παρουσιάζονται οι βασικές καθώς και κάποιες πιο προηγμένες δομές που μας επιτρέπουν να μοντελοποιούμε στην OWL 2.

Κλάσεις και Στιγμιότυπα

Οι *κλάσεις* αντιπροσωπεύουν σύνολα ατόμων. Η δήλωση μιας κλάσης είναι της μορφής:

$$\text{ClassAssertion}(C\ a)$$

όπου C είναι μια έκφραση κλάσης (class expression – κλάσεις για απλότητα) και a είναι ένα όνομα ατόμου. Αυτή η δήλωση αναφέρεται σε ένα άτομο a και δηλώνει ότι αυτό το άτομο έχει τον τύπο C. Σημασιολογικά, αυτό σημαίνει πως το άτομο που έχει το όνομα a βρίσκεται στην επέκταση του συνόλου που περιγράφεται από το C:

$$a \in C$$

Ενσωματωμένες κλάσεις:

- Το σύνολο όλων των ατόμων:

  owl:Thing

- Το κενό σύνολο:

  owl:Nothing

Ιεραρχία Κλάσεων

Η *υπαγωγή κλάσεων* (class subsumption) είναι ένα αξίωμα της μορφής

$$\text{SubClassOf}(C_1\ C_2)$$

όπου C_1 (που δηλώνει την υποκλάση – subclass) και C_2 (που δηλώνει την υπερκλάση – superclass) είναι κλάσεις. Αυτή η δήλωση καταδεικνύει πως η έκφραση C_1 είναι μια υποκλάση της έκφρασης C_2. Σημασιολογικά, αυτό σημαίνει πως κάθε άτομο a το οποίο βρίσκεται στην επέκταση της C_1 βρίσκεται και στην επέκταση της C_2:

$$C_1 \sqsubseteq C_2$$

$$(a \in C_1 \rightarrow a \in C_2)$$

Η *ισότητα κλάσεων* (class equivalence) είναι ένα αξίωμα που καταδεικνύει πως δύο ή περισσότερες κλάσεις αναφέρονται στην ίδια κλάση και είναι της μορφής

$$\text{EquivalentClasses}(C_1\ C_2\ \ldots\ C_n)$$

όπου $n \geq 1$ και C_i με $i = 1, 2, \ldots, n$ είναι κλάσεις.

Ξένες Κλάσεις

Η *ασυμφωνία κλάσεων* (class disjointness) είναι ένα αξίωμα που καταδυκνύει πως δύο ή περισσότερες κλάσεις αναφέρονται σε διαφορετικές κλάσεις, οπότε δεν έχουν κοινά άτομα και είναι της μορφής

$$\text{DisjointClasses}(C_1\ C_2\ \ldots\ C_n)$$

όπου $n \geq 1$ και C_i με $i = 1, 2, \ldots, n$ είναι κλάσεις.

Ιδιότητες Αντικειμένων

Οι ιδιότητες περιγράφουν με ποιόν τρόπο σχετίζονται τα άτομα μεταξύ τους. Μια *ιδιότητα αντικειμένου* (object property) είναι της μορφής

ObjectPropertyAssertion($R\ a\ b$)

όπου a και b είναι ονόματα ατόμων και R είναι μια έκφραση ιδιότητας αντικειμένου (object property expression – ιδιότητα αντικειμένου για απλότητα). Αυτή η δήλωση αναφέρεται σε ένα ρόλο R που συνδέει δύο άτομα a και b. Σημασιολογικά, αυτό σημαίνει πως η πλειάδα των δύο ατόμων (a, b) βρίσκεται στην επέκταση του συνόλου που περιγράφεται από το R:

$$(a, b) \in R$$

Στην OWL2, υπάρχουν οι αρνητικές σχέσεις που περιγράφουν πως μια συγκεκριμένη σχέση μεταξύ δυο ατόμων δεν ισχύει. Μια δήλωση αρνητικής ιδιότητας αντικειμένου είναι της μορφής

NegativeObjectPropertyAssertion($R\ a\ b$)

όπου a και b είναι ονόματα ατόμων και R είναι ιδιότητα αντικειμένου. Αυτή η δήλωση αναφέρεται σε ένα ρόλο R που δεν συνδέει δύο άτομα a και b. Σημασιολογικά, αυτό σημαίνει πως η πλειάδα των δύο ατόμων (a, b) δεν βρίσκεται στην επέκταση του συνόλου που περιγράφεται από το R:

$$(a, b) \notin R$$

Ενσωματωμένες ιδιότητες:

- Η ιδιότητα που συνδέει όλα τα πιθανά ζευγάρια ατόμων:

  owl:topObjectProperty

- Η ιδιότητα που δεν συνδέει κανένα ζευγάρι ατόμων:

  owl:bottomObjectProperty

Ιεραρχία Ιδιοτήτων

Όμοια με την υπαγωγή κλάσεων, η *υπαγωγή ρόλων* (role subsumption) είναι ένα αξίωμα ιδιοτήτων της μορφής

SubObjectPropertyOf($R_1\ R_2$)

όπου R_1 και R_2 είναι ιδιότητες αντικειμένων. Αυτή η δήλωση καταδεικνύει πως ο ρόλος R_1 είναι ένας υπορόλος (subrole) του ρόλου R_2. Σημασιολογικά, αυτό σημαίνει πως κάθε πλειάδα δύο ατόμων (a, b) η οποία βρίσκεται στην επέκταση του συνόλου που περιγράφεται από το R_1 βρίσκεται επίσης και στην επέκταση του συνόλου που περιγράφεται από το R_2, όπου a και b είναι ονόματα ατόμων:

$$R_1 \sqsubseteq R_2$$

$$((a, b) \in R_1 \rightarrow (a, b) \in R_2)$$

Περιορισμοί Πεδίου Ορισμού και Πεδίου Τιμών

Δύο άτομα που συνδέονται μεταξύ τους με μία ορισμένη ιδιότητα αντικειμένου υποδηλώνουν πρόσθετες πληροφορίες για τα ίδια τα άτομα. Τα αξιώματα συμμετοχής στην κλάση είναι της μορφής

ObjectPropertyDomain(R C_1)
ObjectPropertyRange(R C_2)

όπου R είναι μια ιδιότητα αντικειμένου και C_1, C_2 είναι κλάσεις. Η πρώτη δήλωση υποδηλώνει πως εάν ένας ρόλος R συνδέει δύο άτομα a και b, τότε το a είναι τύπου C_1 (πεδίο ορισμού – domain), ενώ η δεύτερη δήλωση υποδηλώνει πως το b είναι τύπου C_2 (πεδίο τιμών – range), όπου a και b είναι ονόματα ατόμων. Σημασιολογικά, αυτό σημαίνει πως η πλειάδα των δύο ατόμων (a, b) βρίσκεται στην επέκταση του συνόλου που περιγράφεται από το R, όπου το άτομα με το όνομα a βρίσκεται στην επέκταση του συνόλου που περιγράφεται από το C_1 και το άτομο με το όνομα b βρίσκεται στην επέκταση του συνόλου που περιγράφεται από το C_2:

$$(a, b) \in R \rightarrow a \in C_1 \wedge b \in C_2$$

Ισότητα και Ανισότητα των Ατόμων

Η *ισότητα των ατόμων* (individual equality) είναι ένα αξίωμα που υποδηλώνει πως δύο ή περισσότερα ονόματα ατόμων αναφέρονται στο ίδιο άτομο και είναι της μορφής

SameIndividual(a_1 a_2 … a_n)

όπου $n \geq 2$, και a_i με $i = 1, 2, \ldots, n$ είναι ονόματα ατόμων.

Αφ' ετέρου, η *ανισότητα των ατόμων* (individual inequality) είναι ένα αξίωμα που υποδηλώνει που δύο ή περισσότερα ονόματα ατόμων δεν αναφέρονται στο ίδιο άτομο και είναι της μορφής

DifferentIndividuals(a_1 a_2 … a_n)

όπου $n \geq 2$ και a_i με $i = 1, 2, \ldots, n$ είναι ονόματα ατόμων.

Να σημειωθεί πως στην OWL η υπόθεση μοναδικού ονόματος δεν ισχύει, και επομένως η OWL δεν κάνει υποθέσεις σχετικά με την ισότητα ή την ανισότητα δύο ατόμων που αναφέρονται με διαφορετικά ονόματα.

Ιδιότητες Τύπου Δεδομένων

Μια δήλωση *ιδιότητας τύπου δεδομένων* (datatype property) έχει σχεδόν την ίδια μορφή με την δήλωση ιδιότητας αντικειμένου

$$\text{DataPropertyAssertion}(R\ a\ v)$$

όπου a είναι ένα όνομα ατόμου, R είναι μια έκφραση ιδιότητας τύπου δεδομένων (datatype property expression – ιδιότητα τύπου δεδομένων για απλότητα) και v είναι ένα λεκτικό (literal). Τα λεκτικά αντιπροσωπεύουν τιμές δεδομένων, όπως συγκεκριμένες ακολουθίες χαρακτήρων (strings) ή ακεραίους. Αυτή η δήλωση αναφέρεται σε ένα ρόλο R που συνδέει ένα άτομο a με μία τιμή δεδομένων (data value) v.

Μια αρνητική δήλωση χρησιμοποιεί την ακόλουθη μορφή αντιστοίχως

$$\text{NegativeDataPropertyAssertion}(R\ a\ v)$$

όπου a είναι ένα όνομα ατόμου, R είναι μια ιδιότητα τύπου δεδομένων και v είναι ένα λεκτικό. Αυτή η δήλωση αναφέρεται σε ένα ρόλο R ο οποίος δεν συνδέει ένα άτομο a με μία τιμή δεδομένων v.

Ενσωματωμένες ιδιότητες:

- Η ιδιότητα που συνδέει όλα τα άτομα με όλα τα λεκτικά:

  owl:topDataProperty

- Η ιδιότητα που δεν συνδέει κανένα άτομο με κάποιο λεκτικό:

  owl:bottomDataProperty

Ένα άτομο που συνδέεται με μία τιμή δεδομένων μέσω μίας συγκεκριμένης ιδιότητας τύπου δεδομένων υποδηλώνει πρόσθετες πληροφορίες σχετικά με το άτομο και την τιμή δεδομένων. Τα αξιώματα συμμετοχής στην κλάση είναι της μορφής

$$\text{DataPropertyDomain}(R\ C)$$
$$\text{DataPropertyRange}(R\ D)$$

όπου R είναι μια ιδιότητα τύπου δεδομένων, C είναι μια κλάση και D είναι ένας τύπος δεδομένων. Η πρώτη δήλωση υποδηλώνει πως εάν ένας ρόλος R συνδέει ένα άτομο a με μία τιμή δεδομένων v, τότε το a έχει τον τύπο C, ενώ η δεύτερη δήλωση υποδηλώνει πως η v είναι τύπου δεδομένων D.

Να σημειωθεί πως ούτε οι δηλώσεις πεδίου ορισμού αλλά ούτε οι δηλώσεις πεδίου τιμών χρησιμοποιούνται σαν περιορισμοί στην γνώση, απλώς επιτρέπουν στον μηχανισμό εξαγωγής συμπερασμάτων να συμπεράνει περαιτέρω γνώση.

Χαρακτηριστικά Ιδιοτήτων

Νέες ιδιότητες αντικειμένων μπορούν να ληφθούν αλλάζοντας την κατεύθυνση ιδιοτήτων αντικειμένων που έχουν δηλωθεί προηγουμένως και ο ορισμός είναι της μορφής

$$\text{InverseObjectProperties}(R_1\ R_2)$$

όπου R_1, R_2 είναι ιδιότητες αντικειμένων. Αυτή η δήλωση υποδηλώνει πως ο ρόλος R_1 είναι ο αντίστροφος ρόλος του ρόλου R_2. Σημασιολογικά, αυτό σημαίνει πως κάθε πλειάδα δύο ατόμων (a, b) που βρίσκεται στην επέκταση του συνόλου που περιγράφεται από τον R_1 υπάρχει επιπλέον μια πλειάδα (b, a) που βρίσκεται στην επέκταση του συνόλου που περιγράφεται από τον R_2, όπου a και b είναι ονόματα ατόμων:

$$((a, b) \in R_1 \rightarrow (b, a) \in R_2)$$

OWL 2 DL & OWL 2 Full

Η έννοια “OWL 2 DL” αναφέρεται στις οντολογίες σε OWL 2 που πληρούν τις προϋποθέσεις για ειδικές συνθήκες και χρησιμοποιούν άμεση σημασιολογία, ενώ η έννοια “OWL 2 Full” αναφέρεται σε σημασιολογία βασισμένη σε RDF. Οι βασικές διαφορές είναι οι ακόλουθες:

- Η OWL 2 DL είναι μια συντακτικά περιορισμένη έκδοση της OWL 2 Full
- Η OWL 2 Full δεν μπορεί να αποφανθεί (undecidable) ενώ η OWL 2 DL μπορεί
- Η OWL 2 DL είναι πλήρως καλυμμένη από διάφορους ποιοτικούς μηχανισμούς εξαγωγής συμπερασμάτων
- Η OWL 2 Full είναι μια επέκταση της RDFS και συνεπώς, ακολουθεί την RDFS σημασιολογία και την γενική συντακτική φιλοσοφία

Η παρούσα διπλωματική είναι βασισμένη στην OWL 2 DL, και ως εκ τούτου είναι βέβαιο πως το δεσμευμένο λεξιλόγιο χρησιμοποιείται μόνο για τον σκοπό που προορίζεται, ισχύουν αυστηρές προϋποθέσεις κατηγοριοποίησης και η δήλωση κλάσεων, τύπων δεδομένων και ιδιοτήτων είναι υποχρεωτική.

2.2. Μηχανική Μάθηση

Σε αυτή την ενότητα παρέχεται μια εισαγωγή στη Μηχανική Μάθηση και παρουσιάζεται το σύστημα Weka που χρησιμοποιήθηκε για την εφαρμογή τεχνικών μηχανικής μάθησης. Επιπλέον, οι βασικές τεχνικές ταξινόμησης, καθώς και οι κανόνες συσχέτισης που χρησιμοποιήθηκαν απεικονίζονται, μαζί με τους δείκτες που καταδεικνύουν την ποιότητα της προκύπτουσας ταξινόμησης ή του κανόνα.

2.2.1. Μάθηση με Επίβλεψη Vs. Μάθηση χωρίς Επίβλεψη

Οι αλγόριθμοι Μηχανικής Μάθησης χωρίζονται σε δύο μεγάλα υποσύνολα, αλγόριθμοι *μάθησης με επίβλεψη* και *χωρίς επίβλεψη*. Στην μάθηση με επίβλεψη, ένα σύνολο δεδομένων εκπαίδευσης εισόδου και οι σχετικές έξοδοι δίνονται στο πρόγραμμα έτσι ώστε να του διδάξουν πως μια συγκεκριμένη είσοδος οδηγεί σε μια συγκεκριμένη έξοδο. Το πρόγραμμα αναλύει τα δεδομένα εκπαίδευσης και παράγει μια συνάρτηση συνεπαγωγής, η οποία μπορεί αργότερα να εφαρμοστεί σε νέα δεδομένα, προκειμένου να χαρτογραφήσει νέες περιπτώσεις. Από την άλλη, στην μάθηση χωρίς επίβλεψη, το πρόγραμμα δεν έχει κάποιο στοιχείο σχετικά με το πώς θα πρέπει να μοιάζει η επιθυμητή έξοδος, καθώς δεν δίνεται σύνολο δεδομένων εκπαίδευσης, και επομένως ψάχνει για συσχετίσεις ανάμεσα στα δεδομένα ώστε να αποκαλύψει μια κρυμμένη δομή σε μη επισημασμένα δεδομένα.

Για τον σκοπό του εμπλουτισμού οντολογιών, επιστρατεύτηκε μάθηση με επίβλεψη. Με βάση μία αρχική οντολογία με ρητές δηλώσεις και ισχυρισμούς, ο στόχος είναι να ταξινομηθεί η οντολογία προς εμπλουτισμό, ώστε να επισημανθούν προηγουμένως μη επισημασμένα δεδομένα καθώς και να παραχθούν κανόνες συσχέτισης.

Ταξινόμηση

Ταξινόμηση είναι η διαδικασία της επισήμανσης στοιχείων, εντοπίζοντας σε ποια κατηγορία από ορισμένο σύνολο κατηγοριών ανήκει, με βάση ένα σύνολο από προηγουμένως επισημασμένων παρατηρήσεων. Κάθε παρατήρηση, που ονομάζεται στιγμιότυπο, αναλύεται σε ένα σύνολο από μετρήσιμες ιδιότητες που ονομάζονται επεξηγηματικές μεταβλητές ή χαρακτηριστικά. Ο προκύπτων αλγόριθμος που πραγματοποιεί την ταξινόμηση, με την ονομασία ταξινομητής (classifier), χρησιμοποιείται για την εκχώρηση ετικετών στο σύνολο δεδομένων δοκιμής, όπου οι τιμές των χαρακτηριστικών που χρησιμοποιούνται στην πρόγνωση είναι γνωστές, αλλά η τιμή της ετικέτας ταξινόμησης είναι άγνωστη.

Μερικές από τις αμέτρητες εφαρμογές της ταξινόμησης είναι οι ακόλουθες:

- Όραση υπολογιστών
- Αναγνώριση φωνής
- Αναγνώριση γραφικού χαρακτήρα
- Επεξεργασία φυσικής γλώσσας
- Ταξινόμηση εγγράφων
- Μηχανές αναζήτησης στο Διαδίκτυο
- Αναγνώριση προτύπων

Δεν υπάρχει ένας καθολικά καλύτερος ταξινομητής, η απόδοση εξαρτάται κυρίως από τα χαρακτηριστικά των δεδομένων που χρήζουν ταξινόμησης. Η απόδοση και η

ποιότητα ενός ταξινομητή αξιολογούνται από δείκτες, μερικοί από τους οποίους παρουσιάζονται ακολούθως.

Η *ακρίβεια* (precision) και η *ανάκληση* (recall or sensitivity) είναι μετρήσεις της σχετικότητας. Η ακρίβεια επικεντρώνεται στη χρησιμότητα των αποτελεσμάτων, ενώ η ανάκληση αναφέρεται στο πόσο πλήρη είναι τα αποτελέσματα. Κατά την ταξινόμηση, η ακρίβεια είναι ο αριθμός των αληθώς θετικών (true positives[1]) διαιρεμένος με τον συνολικό αριθμό των στοιχείων που έχουν επισημανθεί πως ανήκουν στην θετική κλάση (αληθώς θετικά – true positives και ψευδώς θετικά – false positives[2]), οπότε η χαμηλή ακρίβεια υποδεικνύει μεγάλο αριθμό ψευδώς θετικών. Αφ' ετέρου, η ανάκληση είναι ο αριθμός των αληθώς θετικών διαιρεμένος με τον συνολικό αριθμό των στοιχείων που όντως ανήκουν στην θετική κλάση (αληθώς θετικά – true positives και ψευδώς αρνητικά – false negatives[3]), οπότε η χαμηλή ανάκληση υποδεικνύει μεγάλο αριθμό ψευδώς αρνητικών.

Η ακρίβεια υποδεικνύει το ποσοστό των σωστών αποτελεσμάτων (true positives και true negatives[4]) ανάμεσα στο συνολικό αριθμό των περιπτώσεων που εξετάστηκαν.

Στην παρούσα διπλωματική γίνεται χρήση δύο αλγορίθμων ταξινόμησης: **RandomForest Tree Classifier** [7] & **kNN (IBk) Lazy Classifier** [8]. Ο RandomForest ταξινομητής δημιουργεί ένα «δάσος» από RandomTrees, όπου τα στοιχεία που χρησιμοποιούνται για τη δημιουργία των επιμέρους δέντρων απόφασης (decision trees) έχουν επιλεγεί τυχαία. Η τελική απόφαση για την ταξινόμηση του αντικειμένου προκύπτει από την συνηθέστερη απόφαση των επιμέρους δέντρων. Κατά την ταξινόμηση με τον kNN (k-Nearest Neighbours) ταξινομητή, ένα αντικείμενο ταξινομείται με βάση την πλειονότητα των γειτόνων του, όπου το αντικείμενο εν τέλει ανατίθεται στην κλάση που είναι η συνηθέστερη ανάμεσα στους *k* πιο κοντινούς του γείτονες (το *k* είναι ένας θετικός ακέραιος, συνηθίζεται μικρός). Εάν *k* = 1, τότε το αντικείμενο ανατίθεται στην κλάση του κοντινότερου γείτονα.

Κανόνες Συσχέτισης

Η *εκμάθηση κανόνων συσχέτισης* είναι μια διαδικασία ανακάλυψης σχέσεων μεταξύ μεταβλητών σε ένα σύνολο δεδομένων. Εάν ένα σύνολο δεδομένων αποτελείται από n διαφορετικά χαρακτηριστικά/στοιχεία $I = \{i_1, i_2, \dots, i_n\}$ και m διαφορετικές καταγραφές/συναλλαγές $T = \{t_1, t_2, \dots, t_m\}$, όπου κάθε καταγραφή υποδηλώνει με αληθές ή ψευδές εάν περιλαμβάνει κάθε στοιχείο και, ως αποτέλεσμα, κάθε συναλλαγή στο T έχει ένα μοναδικό ID και περιλαμβάνει ένα υποσύνολο των στοιχείων του I, τότε ένας κανόνας υποδεικνύει πως

$$X \Rightarrow Y$$

[1] True positives: στοιχεία τα οποία είναι σωστά επισημασμένα πως ανήκουν στην θετική κλάση.
[2] False positives: στοιχεία τα οποία είναι λάθος επισημασμένα πως ανήκουν στην θετική κλάση.
[3] False negatives: στοιχεία τα οποία ανήκουν στην θετική κλάση αλλά δεν επισημάνθηκαν.
[4] True negatives: στοιχεία τα οποία δεν ανήκουν στην θετική κλάση και ορθώς δεν επισημάνθηκαν.

όπου $X, Y \subseteq I$ και $X \cap Y = \emptyset$. Αυτό σημαίνει πως εάν κάθε χαρακτηριστικό που αντιστοιχεί σε ένα στοιχείο στο σύνολο X (antecedent) είναι αληθές σε μια συναλλαγή, τότε τα χαρακτηριστικά που αντιστοιχούν στα στοιχεία του συνόλου Y (consequent) είναι εξίσου αληθή.

Η παραγωγή ικανοποιητικών κανόνων συσχέτισης εξαρτάται κυρίως από την εφαρμογή ενός ελάχιστου ορίου στήριξης (minimum support threshold) για την εύρεση όλων των συχνών συνόλων στοιχείων σε ένα σύνολο δεδομένων και ένα ελάχιστο όριο εμπιστοσύνης (minimum confidence constraint) για τον σχηματισμό κανόνων στα συχνά σύνολα στοιχείων. Η απόδοση ενός κανόνα μετράται με βάση τις παρακάτω μετρικές (metrics):

- Support
- Confidence
- Lift
- Conviction

Support ενός συνόλου στοιχείων X

$$supp(X)$$

σε σχέση με ένα σύνολο συναλλαγών T, είναι το ποσοστό των συναλλαγών στο σύνολο δεδομένων που περιέχουν το σύνολο στοιχείων X.

Confidence ενός κανόνα $X \Rightarrow Y$

$$conf(X \Rightarrow Y)$$

σε σχέση με ένα σύνολο συναλλαγών T, είναι το ποσοστό των συναλλαγών που περιέχουν τα X και Y ταυτόχρονα. Υπολογίζεται ως ακολούθως:

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

Lift ενός κανόνα $X \Rightarrow Y$

$$lift(X \Rightarrow Y)$$

είναι η αναλογία του support που παρατηρήθηκε προς του προσδοκώμενου εάν τα X και Y ήταν ανεξάρτητα:

$$lift(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)}$$

Τέλος, *conviction* ενός κανόνα $X \Rightarrow Y$ ορίζεται ως ακολούθως:

$$conv(X \Rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \Rightarrow Y)}$$

Στην παρούσα διπλωματική γίνεται χρήση του αλγορίθμου εκμάθησης κανόνων συσχέτισης Apriori [9]. Ο αλγόριθμος αναγνωρίζει τα αντικείμενα που εμφανίζονται συχνά σε ένα σύνολο δεδομένων και τα επεκτείνει σε όλο και μεγαλύτερα σύνολα

στοιχείων μέχρις ότου τα σύνολα που προκύπτουν να εμφανίζονται αρκετά συχνά στα δεδομένα συναλλαγών με βάση το όριο της εμπιστοσύνης που έχει θέσει ο χρήστης.

2.2.2. Weka Software

Το **Weka** [10] είναι ένα ελεύθερο λογισμικό που έχει εκδοθεί υπό την GNU General Public License[5] άδεια. Πρόκειται για ένα λογισμικό που παρέχει μια συλλογή από έτοιμους προς χρήση αλγόριθμους μηχανικής μάθησης για εργασίες εξόρυξης δεδομένων. Οι αλγόριθμοι μπορούν είτε να εφαρμοστούν απευθείας σε ένα σύνολο δεδομένων, είτε, όπως γίνεται και στην παρούσα διπλωματική, να κληθούν από ένα κώδικα Java. Το Weka περιέχει εργαλεία για προεπεξεργασία δεδομένων, ταξινόμηση, παλινδρόμηση, ομαδοποίηση, κανόνες συσχέτισης και απεικόνιση δεδομένων.

Το Weka δέχεται ως είσοδο αρχεία ARFF. Ένα ARFF (Attribute-Relation File Format) αρχείο είναι ένα αρχείο κειμένου ASCII που περιγράφει μια λίστα από στιγμιότυπα που μοιράζονται ένα σύνολο χαρακτηριστικών (attributes). Αποτελείται από δύο διακριτές ενότητες: *πληροφορίες επικεφαλίδας* (Header Information) και *πληροφορίες δεδομένων* (Data Information).

Η επικεφαλίδα του αρχείου ARFF περιέχει το όνομα της σχέσης, μία λίστα από χαρακτηριστικά (τις στήλες στα δεδομένα), και τους τύπους τους. Τα δεδομένα του αρχείου ARFF περιέχουν τα στιγμιότυπα δεδομένων, όπου κάθε στιγμιότυπο αναπαρίσταται σε μια γραμμή και οι τιμές των χαρακτηριστικών για κάθε στιγμιότυπο οριοθετούνται από κόμμα.

Τα *χαρακτηριστικά* (attributes) δηλώνονται μοναδικά με το όνομα και τον τύπο δεδομένων. Στο Weka υποστηρίζονται οι ακόλουθοι τύποι δεδομένων:

- Numeric - Αριθμητικά:

  Τα αριθμητικά attributes είναι πραγματικοί ή ακέραιοι αριθμοί και ο ορισμός τους είναι:

  ```
  @ATTRIBUTE num     numeric
  ```

- Nominal - Ονομαστικά:

  Τα ονομαστικά attributes είναι attributes που δηλώνουν μια συγκεκριμένη λίστα με όλες τις πιθανές τιμές στον ορισμό τους:

  ```
  @ATTRIBUTE nom     {nom_val1, nom_val2, nom_val3}
  ```

[5] http://www.gnu.org/licenses/gpl.html

- String - Ακολουθίες Χαρακτήρων:

  Τα attributes ακολουθίας χαρακτήρων περιλαμβάνουν αυθαίρετες τιμές κειμένου και ο ορισμός τους είναι:

  @ATTRIBUTE str string

- Date - Ημερομηνίες:

  Τα attributes ημερομηνίας περιέχουν τιμές ημερομηνίας και, αν ορισθεί, με συγκεκριμένη μορφοποίηση ημερομηνίας (προαιρετικά):

  @ATTRIBUTE dat date [<date-format>]

  Η προεπιλεγμένη μορφοποίηση βασίζεται στη μορφή ISO-8601 συνδυασμού ημερομηνίας και ώρας:

  yyyy-MM-dd HH:mm:ss

Τέλος, οι τιμές των attributes πρέπει να εμφανίζονται με τη σειρά με την οποία έχουν δηλωθεί στην επικεφαλίδα και εάν μια τιμή απουσιάζει, τότε εκπροσωπείται από ένα μόνο ερωτηματικό (Missing Value – ?).

Κεφάλαιο 3. Σχετικές Εργασίες

Η σημασία του εμπλουτισμού οντολογιών αντανακλάται στο μεγάλο αριθμό εργασιών που έχουν ασχοληθεί με αυτό το πρόβλημα. Στο παρόν κεφάλαιο παρουσιάζονται κάποιες από τις μεθόδους που βοήθησαν στην σύλληψη της ιδέας του προτεινόμενου μοντέλου.

3.1. WordNet

Το WordNet [11] είναι μια μεγάλη λεξικογραφική βάση δεδομένων στα Αγγλικά που επιφανειακά μοιάζει με θησαυρό όρων (thesaurus), στο ότι συγκεντρώνει λέξεις με βάση τη σημασία τους. Ωστόσο, υπάρχουν ορισμένες σημαντικές διακρίσεις. Κατ' αρχάς, το WordNet διασυνδέει όχι μόνο μορφές λέξεων αλλά συγκεκριμένες σημασίες λέξεων. Κατά δεύτερον, το WordNet επισημαίνει τις σημασιολογικές σχέσεις ανάμεσα στις λέξεις, ενώ οι ομαδοποιήσεις λέξεων σε ένα θησαυρό όρων δεν ακολουθεί κάποιο ρητό μοτίβο εκτός από την σημασιολογική ομοιότητα. Η δομή του WordNet το κάνει χρήσιμο εργαλείο στην υπολογιστική γλωσσολογία και την επεξεργασία φυσικής γλώσσας.

3.2. REEL

Ένα σύστημα εξαγωγής σχέσεων (Relationship Extraction System) ανιχνεύει εάν υπάρχει σημασιολογική σχέση, συνήθως ενός ορισμένου τύπου, ανάμεσα σε δύο ή περισσότερες έννοιες εντός ενός πλαισίου όπου και εμφανίζονται. Το REEL (RElationship Extraction Learning Framework) [12] είναι ένα framework εξαγωγής σχέσεων για Java που επιτρέπει στους χρήστες να εφαρμόσουν διαφορετικά συστήματα εξαγωγής σχέσεων σε μερικά απλά βήματα. Το REEL είναι κυρίως μια προσπάθεια να έρθουν διαφορετικά συστήματα και εργαλεία εξαγωγής σχέσεων σε ένα ελεγχόμενο και ενοποιημένο περιβάλλον.

3.3. NELL

Το NELL (Never-Ending Language Learner) [13] είναι ένα υπολογιστικό σύστημα που μαθαίνει με την πάροδο του χρόνου να διαβάζει το διαδίκτυο. Ο στόχος του ερευνητικού προγράμματος "Read the Web" at Carnegie Mellon University είναι η οικοδόμηση ενός συστήματος ατέρμονης μηχανική μάθησης που αποκτά την ικανότητα να εξάγει δομημένες πληροφορίες από μη δομημένες ιστοσελίδες. Εάν επιτύχει, αυτό θα οδηγήσει σε μια βάση γνώσης (μια σχεσιακή βάση δεδομένων) από δομημένες πληροφορίες που αντικατοπτρίζει το περιεχόμενο του διαδικτύου. Με μια αρχική οντολογία που καθορίζει εκατοντάδες κατηγορίες και σχέσεις, καθώς επίσης και 10 με 15 παραδείγματα σε κάθε κατηγορία και ρόλο, και επιπλέον μια συλλογή από ιστοσελίδες σαν είσοδο, το NELL τρέχει 24 ώρες την ημέρα, συνεχόμενα για να εξάγει νέα στιγμιότυπα των κατηγοριών και των σχέσεων και για να μάθει "better than the day before".

3.4. SOFIE

Ένα σύστημα χωρίς επίβλεψη για αυτοματοποιημένο εμπλουτισμό οντολογιών είναι το σύστημα SOFIE [14]. Αναλύει έγγραφα φυσικής γλώσσας, εξάγει οντολογικά γεγονότα από αυτά και συντάσσει αυτά τα νέα γεγονότα στην οντολογία. Οι λέξεις αποσαφηνίζονται στην πιο πιθανή τους σημασία με εφαρμογή λογικών μηχανισμών εξαγωγής συμπερασμάτων στην υπάρχουσα γνώση. Καθώς τα πρόσφατα εξαγόμενα γεγονότα πρέπει να είναι συνεπή με την υπάρχουσα οντολογία, ο αλγόριθμος Weighted MAX-SAT χρησιμοποιείται για να αντιμετωπίσει το πρόβλημα. Τα αποτελέσματα από διαφορετικά σώματα έδειξαν έως και 94.7 % precision και 31.08 % recall.

3.5. OTTO

Ένα άλλο framework, το OTTO [15] επιτρέπει την ημι-αυτόματη κατασκευή ή εμπλουτισμό οντολογιών. Στο συγκεκριμένο σύστημα εκτελούνται τρεις αλγόριθμοι. Πρώτα, ένας αλγόριθμος εννοιολογικής ομαδοποίησης χρησιμοποιήθηκε για την κατασκευή ταξινομιών (taxonomies). Στη συνέχεια οι ταξινομίες κατασκευάζονται με συνδυασμό πληροφοριών από τα μοτίβα WordNet και Hearst. Τέλος, μη-ταξινομικές σχέσεις καθορίζονται από την εξαγωγή συντακτικών πλαισίων των όρων εισόδου. Στον τομέα της ομαδοποίησης κειμένου και ταξινόμησης με το WordNet ως οντολογία, η μέθοδος επιτυγχάνει αρκετές βελτιώσεις.

3.6. Συμπεράσματα

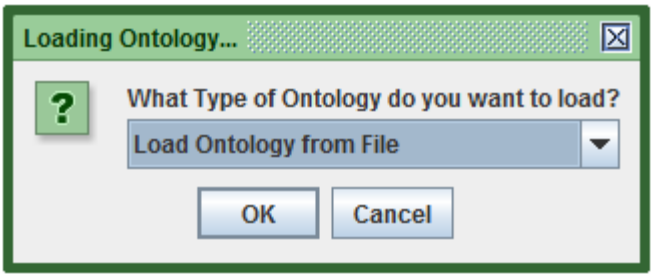
Το πρόβλημα του εμπλουτισμού οντολογιών είναι ακόμα ανοιχτό και οι μέθοδοι εκμάθησης οντολογιών είναι σχεδόν αδύνατο να εφαρμοστούν στην πράξη. Στο επόμενο κεφάλαιο, ορίζεται η δική μας προσέγγιση, η οποία επικεντρώνεται στον εμπλουτισμό οντολογιών εξαρτώμενο από μια αρχική οντολογία που παρέχει ο χρήστης και της οποίας τα αξιώματα μπορεί να είναι εφαρμόσιμα και χρήσιμα για τον σκοπό του εμπλουτισμού.

Κεφάλαιο 4. Προτεινόμενο Μοντέλο

Σε αυτό το κεφάλαιο, παρουσιάζεται το προτεινόμενο μοντέλο της παρούσας διπλωματικής μαζί με βασικά σημεία της υλοποίησης του κώδικα που παράχθηκε ώστε να πραγματοποιηθεί το παρόν μοντέλο.

4.1. Φόρτωση Οντολογιών

Η πρώτη εργασία που εκτελεί το πρόγραμμα είναι να φορτώνει τις οντολογίες. Η πρώτη οντολογία που δέχεται είναι αυτή που παρέχει ο χρήστης ως πλήρη με ρητές δηλώσεις που θα χρησιμοποιηθεί για την εκπαίδευση των ταξινομητών. Αρχικά, εμφανίζεται ένα παράθυρο για να δώσει στον χρήστη τις επιλογές να φορτώσει μια οντολογία που να προέρχεται από ένα αρχείο στον υπολογιστή του χρήστη ή να δώσει ένα URI και να φορτωθεί η οντολογία από το διαδίκτυο. Ανάλογα με την επιλογή του χρήστη, η οντολογία φορτώνεται στο πρόγραμμα.



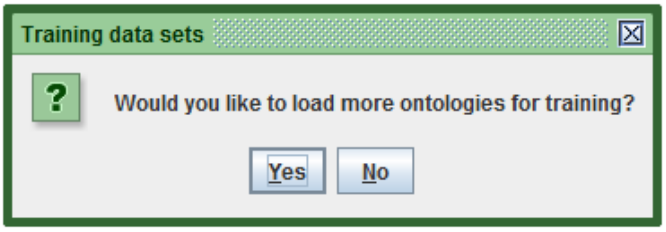
Εικόνα 1 Φόρτωση Οντολογιών: Οι πιθανές επιλογές είναι "Load Ontology from File" και "Load Ontology from the Web"

Στη συνέχεια, ο χρήστης εισάγει αν οι σχέσεις θα αποθηκεύονται ως υπαρξιακές. Αυτό αναφέρεται στις ιδιότητες αντικειμένου της οντολογίας και την μετατροπή από το αρχείο OWL σε ένα αρχείο ARFF. Με την αποθήκευση των σχέσεων ως υπαρξιακές, τα χαρακτηριστικά (attributes) που αναφέρονται σε ιδιότητες αντικειμένου θα δηλώνουν μόνο πως ένας συγκεκριμένος ισχυρισμός ρόλου ισχύει για ένα άτομο αλλά χωρίς να αναφέρεται το άλλο άτομο που συνδέεται με το πρώτο μέσω του ρόλου.

Εικόνα 2 Ο χρήστης ερωτάται για την αποθήκευση των σχέσεων ως υπαρξιακές ή όχι.

Ένας μηχανισμός εξαγωγής συμπερασμάτων εκκινείται με βάση τη δοσμένη οντολογία για να εξάγει κάθε ισχυρισμό που αφορά κλάσεις, ιδιότητες αντικειμένων και ιδιότητες τύπου δεδομένων. Για κάθε άτομο που αναφέρεται στην οντολογία, εξάγονται όλες οι κλάσεις που ανήκει το άτομο και αποθηκεύονται σε μία δομή HashMap. Ακολούθως, οι ιδιότητες αντικειμένων και τύπου δεδομένων που συμμετέχει το συγκεκριμένο άτομο εξάγονται και αποθηκεύονται. Εάν ο χρήστης ζήτησε να διατηρηθούν οι σχέσεις ως υπαρξιακές, οι ιδιότητες αντικειμένων θα αποθηκευτούν μόνο με το όνομά τους και ως δυαδικά attributes. Ομοίως, εάν ο χρήστης δεν ζήτησε τη διατήρηση των σχέσεων ως υπαρξιακές, οι ιδιότητες θα αποθηκευτούν και πάλι ως δυαδικά attributes, απλώς όχι μόνο με το όνομά τους, αλλά και με το όνομα του ατόμου που η κάθε σχέση συνδέει το αρχικό άτομο. Στην περίπτωση των ιδιοτήτων τύπου δεδομένων, η μορφή που θα αποθηκευτούν εξαρτάται από το πεδίο τιμών της κάθε ιδιότητας. Το πρόγραμμα διαχωρίζει τους προκαθορισμένους τύπους δεδομένων σε τρεις διαφορετικές κατηγορίες τύπων: Δυαδικών (Boolean), Αριθμητικών (Numerical) και Αλφαριθμητικών (String). Ο λόγος αυτού του διαχωρισμού είναι επειδή η μορφή ARFF λειτουργεί μόνο με συγκεκριμένους τύπους δεδομένων, όπως αναφέρθηκε στο Κεφάλαιο 2.

Σε περίπτωση που ο χρήστης επιθυμεί να φορτώσει περισσότερες οντολογίες για να χρησιμοποιηθούν ως σύνολα δεδομένων εκπαίδευσης, το επόμενο παράθυρο εμφανίζεται ώστε να επαναληφθεί ή όχι η διαδικασία της φόρτωσης μια οντολογίας και αποθήκευσης της στην δομή HashMap προτού δημιουργηθεί το επιθυμητό αρχείο ARFF. Εάν ο χρήστης επιλέξει να φορτώσει επιπλέον οντολογίες, η προαναφερθείσα διαδικασία εξαγωγής ατόμων, κλάσεων, ιδιοτήτων αντικειμένων και τύπων δεδομένων επαναλαμβάνεται έως ότου ο χρήστης επιλέξει την επιλογή "**No**".



Εικόνα 3 Ο χρήστης ερωτάται εάν επιθυμεί να εισάγει επιπλέον οντολογίες πριν δημιουργηθεί το συνολικό training dataset.

Μόλις ο χρήστης επιλέξει να μην φορτώσει επιπλέον οντολογίες, καλείται μια άλλη μέθοδος η οποία δημιουργεί το επιθυμητό ARFF αρχείο από τη δομή HashMap που δημιουργήθηκε με τη συλλογή αντικειμένων από όλες τις οντολογίες που παρείχε ο χρήστης. Το πρώτο attribute είναι πάντα τα άτομα (individuals) τα οποία αποθηκεύονται ως strings. Τα επόμενα attributes είναι τα ονόματα των κλάσεων (classes) που αποθηκεύονται ως nominal attributes με πιθανές τιμές "**true**" και "**false**". Στη συνέχεια, οι ιδιότητες αντικειμένων (object properties) και οι αντίστροφές τους αποθηκεύονται εξίσου ως nominal attributes με πιθανές τιμές "**true**" και "**false**". Τέλος, αποθηκεύονται οι ιδιότητες τύπου δεδομένων (data properties) και σε αυτή την περίπτωση ο τύπος του κάθε attribute ποικίλει ανάλογα με τον πεδίο τιμών της κάθε ιδιότητας. Οι ιδιότητες αποθηκεύονται ως numeric attributes εάν το πεδίο τιμών είναι κάποιο από τα ακόλουθα: *decimal, float, double, integer, positiveInteger, nonPositiveInteger, negativeInteger, nonNegativeInteger, long, int, short, byte, unsignedLong, unsignedInt, unsignedShort, unsignedByte, hexBinary, base64Binary*. Αποθηκεύονται ως date attributes εάν το πεδίο τιμών είναι ένα από τα ακόλουθα: *dateTime, time, date, gYearMonth, gYear, gMonthDay, gDay, gMonth*. Στο παρόν μοντέλο ήταν γνωστό πως καμία ιδιότητα τύπου δεδομένων δεν θα έχει ημερομηνία ως πεδίο τιμών και επομένως δεν γίνεται κάποια παραπάνω αναφορά σε αυτό τον τύπο attributes. Τέλος, οι ιδιότητες αποθηκεύονται ως string attributes εάν το πεδίο τιμών δεν είναι ορισμένο ή είναι κάποιο από τα ακόλουθα: *string, normalizedString, token, language, NMTOKEN, Name, NCName, anyURI*.

Η παραπάνω διαδικασία ακολουθείται για τη δημιουργία της επικεφαλίδας (header) του αρχείου ARFF. Το κομμάτι των δεδομένων του αρχείου δημιουργείται όπως περιγράφεται ακολούθως: Αρχικά, το όνομα του ατόμου που αναφέρεται εισάγεται

ως τιμή για το *Individuals* attribute. Στη συνέχεια, για κάθε attribute κλάσης ερευνάται η δομή HashMap για το κατά πόσο το συγκεκριμένο άτομο ανήκει στην κλάση και ανάλογα με τα ευρήματα, η τιμή ορίζεται ως "**true**" ή "**false**". Η ίδια διαδικασία ακολουθείται και για τις ιδιότητες αντικειμένων, όπου ερευνάται κατά πόσο το άτομο συμμετέχει σε ένα συγκεκριμένο ισχυρισμό ρόλου. Για τις ιδιότητες τύπου δεδομένων, ανάλογα με το πεδίο τιμών της ιδιότητας, η τιμή εισάγεται κατά συνέπεια με το όνομα των αντίστοιχων attributes. Εάν δεν υπάρχουν τιμές για να καταχωρηθούν ως τιμή της ιδιότητας για το συγκεκριμένο άτομο, τότε εισάγεται τιμή έλλειψης (Missing Value).

Στη συνέχεια εκτίθεται ένα παράδειγμα του αρχείου ARFF που δημιουργείται ώστε να παρουσιαστεί η μορφή των attributes στην επικεφαλίδα καθώς και το πως κατασκευάζονται οι αντίστοιχες τιμές στην ενότητα των δεδομένων. Για να γίνει εμφανής η διαφορά ανάμεσα στη αποθήκευση των σχέσεων ως υπαρξιακών ή όχι, παρέχονται τα ακόλουθα δείγματα επικεφαλίδας ARFF μετά την εφαρμογή του προγράμματος στην οντολογία οίνου (wine ontology) [16]. Η πρώτη κεφαλίδα προέκυψε μετά την επιλογή υπαρξιακών σχέσεων, ενώ η επόμενη είναι εμφανές πως διατηρεί και τα ονόματα των σχετιζόμενων ατόμων.

Παράδειγμα: Header όπου οι Σχέσεις έχουν διατηρηθεί ως υπαρξιακές

```
@relation wine

@attribute Individuals string
@attribute Anjou_class {true,false}
@attribute Beaujolais_class {true,false}
@attribute CabernetFranc_class {true,false}
@attribute CabernetSauvignon_class {true,false}
@attribute Chardonnay_class {true,false}
…
@attribute WineSugar_class {true,false}
@attribute Winery_class {true,false}
@attribute Zinfandel_class {true,false}
@attribute adjacentRegion_inverseObjectProperty {true,false}
@attribute adjacentRegion_objectProperty {true,false}
@attribute hasBody_inverseObjectProperty {true,false}
@attribute hasBody_objectProperty {true,false}
…
@attribute madeFromGrape_inverseObjectProperty {true,false}
@attribute madeFromGrape_objectProperty {true,false}
@attribute yearValue_dataProperty numeric

@data
AlsaceRegion,false,false,false,false,false,false,false,false,false,false,false,false,fa
lse,false,false,false,false,false,false,false,false,false,true,false,false,false,false,false,false,false,
false,false,false,false,false,false,false,false,false,false,false,false,false,false,false,false,false,fal
se,false,false,false,false,false,false,false,false,false,false,false,false,false,false,false,false,false,
false,false,false,false,false,false,false,false,false,false,false,false,false,false,false,false,false,fal
se,false,false,false,false,false,false,false,false,false,false,false,false,false,false,false,false,false,
```

false,false,false,false,false,false,false,false,false,false,false,false,false,false,true,false,false,false,false,false,false,false,false,false,false,false,false,false,false,?

AnjouRegion,false,true,false,true,false,false,false,false,false,false,false,false,false,false,false,false,?

…

Παράδειγμα: Header όπου οι Σχέσεις διατηρούν την σύνδεση με κάθε διαφορετικό Άτομο

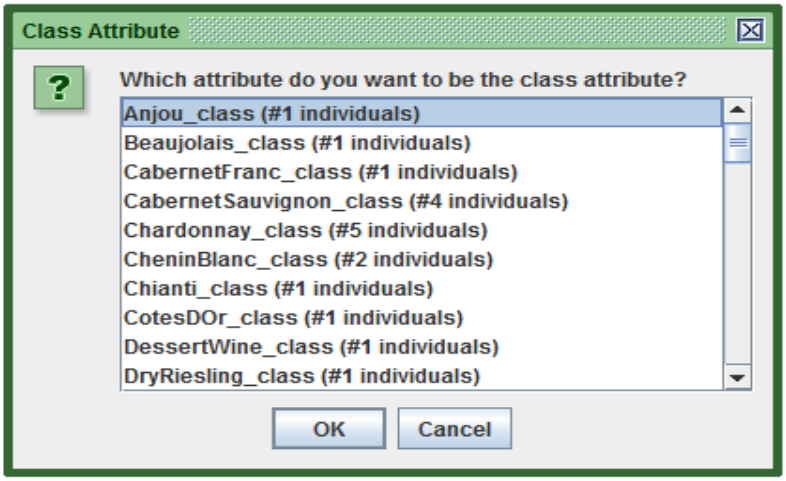
```
@relation wine

@attribute Individuals string
@attribute Anjou_class {true,false}
@attribute Beaujolais_class {true,false}
@attribute CabernetFranc_class {true,false}
@attribute CabernetSauvignon_class {true,false}
@attribute Chardonnay_class {true,false}
…
@attribute WineSugar_class {true,false}
@attribute Winery_class {true,false}
@attribute Zinfandel_class {true,false}
@attribute adjacentRegion_inverseObjectProperty {true,false}
@attribute adjacentRegion_objectProperty_SonomaRegion {true,false}
@attribute hasBody_inverseObjectProperty {true,false}
@attribute hasBody_objectProperty_Full {true,false}
@attribute hasBody_objectProperty_Light {true,false}
@attribute hasBody_objectProperty_Medium {true,false}
…
@attribute madeFromGrape_inverseObjectProperty {true,false}
@attribute madeFromGrape_objectProperty_SauvignonBlancGrape {true,false}
@attribute madeFromGrape_objectProperty_SemillonGrape {true,false}
@attribute yearValue_dataProperty numeric
```

Πριν την αποθήκευση του αρχείου ARFF και τη χρήση τους ως training dataset, κάθε attribute διερευνάται και στην περίπτωση που όλες οι τιμές του είναι ίδιες για όλα τα στιγμιότυπα (instances), τότε το attribute διαγράφεται, καθώς δεν έχει τίποτα να προσφέρει στην διαδικασία της ταξινόμησης.

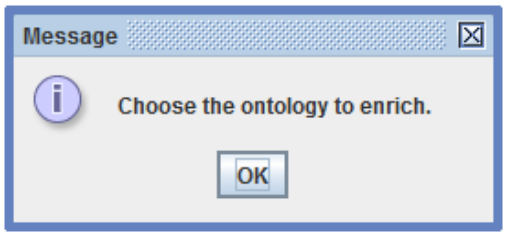
Σαν τελικό βήμα πριν την αποθήκευση του dataset, το επόμενο παράθυρο εμφανίζεται και παρουσιάζει στον χρήστη όλα τα πιθανά attributes που μπορούν να χρησιμοποιηθούν ως class attributes και με βάση τα οποία θα γίνει αργότερα η ταξινόμηση του testing dataset. Η επιλογή γίνεται από μια λίστα με όλα τα attributes

που αντιστοιχούν σε κλάσεις και ιδιότητες αντικειμένων. Παράλληλα με κάθε attribute ο αριθμός των ατόμων που ανήκουν στη σχετική κλάση ή συμμετέχουν στη σχετική ιδιότητα εμφανίζεται εντός μίας παρένθεσης για να διευκολύνει το χρήστη να επιλέξει το πιο κατάλληλο class attribute. Το όνομα του class attribute αποθηκεύεται σε μια καθολική μεταβλητή ώστε να δηλωθεί το class attribute αργότερα στη διαδικασία και στο training αλλά και στο testing dataset.



Εικόνα 4 Ζητείται από τον χρήστη να επιλέξει το class attribute από μία λίστα με όλα τα πιθανά attributes.

Όταν επιλεγεί και το class attribute, το πρώτο αρχείο ARFF που διατηρεί το training dataset αποθηκεύεται με την ονομασία *trainData.arff*. Συνακόλουθα, εμφανίζεται ένα μήνυμα για να ενημερώσει τον χρήστη πως η διαδικασία φόρτωσης που ακολουθεί αφορά την οντολογία που θα χρησιμοποιηθεί ως testing dataset και που χρειάζεται εμπλουτισμό.



Εικόνα 5 Ο χρήστης ενημερώνεται για την φόρτωση την οντολογίας προς εμπλουτισμό.

Το παράθυρο που ακολουθεί, είναι το ίδιο με αυτό της Εικόνας 1, το οποίο ζητά από την χρήστη να επιλέξει αν θα φορτώσει μία οντολογία από αρχείο ή από το διαδίκτυο. Ανάλογα με την επιλογή του χρήστη, η οντολογία φορτώνεται στο πρόγραμμα ακολουθώντας την ίδια ακριβώς διαδικασία με προηγουμένως, όπου κάθε άτομο, κλάση, ιδιότητα αντικειμένων και τύπου δεδομένων εξάγεται και αποθηκεύεται σε

μία δομή HashMap ώστε να δημιουργηθεί το νέο αρχείο ARFF με την ονομασία *testData.arff*.

4.2. Προετοιμασία για την Ταξινόμηση

Το επόμενο βήμα είναι η εξέταση του κατά πόσο τα δύο datasets είναι συμβατά και εάν δεν είναι να επέλθει η κατάλληλη διαδικασία ώστε να γίνουν και να εκπαιδευτεί ο ταξινομητής βασιζόμενος στο πρώτο dataset και να εφαρμοστεί στο δεύτερο. Τα δύο datasets είναι συμβατά εάν έχουν την ίδια επικεφαλίδα και το ίδιο class attribute. Στην παρούσα περίπτωση, το class attribute δεν έχει ακόμα δηλωθεί, καθώς το testing dataset είναι πιθανό να μην περιέχει καν το class attribute. Αυτό είναι λογικό να συμβεί καθώς ο ταξινομητής θα εφαρμοστεί με σκοπό τον εμπλουτισμό της οντολογίας με νέους ισχυρισμούς, κάποιοι από τους οποίους πιθανό να είναι ισχυρισμοί κλάσεων σχετικά με κάποια κλάση που δεν είχε ορισθεί προηγουμένως και ο χρήστης επιλέγει από το training dataset ως πιθανή κλάση για εμπλουτισμό της αρχικής οντολογίας.

Για να γίνουν συμβατά τα δύο datasets, το testing dataset εξετάζεται εξονυχιστικά με έλεγχο του κάθε attribute και εάν δεν ανήκει κάποιο attribute και στο training dataset, τότε διαγράφεται. Επιπλέον, διερευνάται κάθε attribute του training dataset και εάν δεν ανήκει και στο testing dataset, τότε εισάγεται ακριβώς στην ίδια θέση που διατηρείται και στο training dataset. Οι κλάσεις και οι ιδιότητες είχαν ταξινομηθεί πριν τη δημιουργία των datasets, ώστε να επιβεβαιωθεί πως και οι δύο κεφαλίδες θα είναι ταυτόσημες μετά τις προαναφερθείσες επαναλήψεις.

Αφού οι κεφαλίδες γίνουν συμβατές, το class attribute ορίζεται και στα δύο dataset. Στην περίπτωση που το class attribute είναι τύπου string, εφαρμόζεται το φίλτρο *StringToNominal* στο συγκεκριμένο attribute, ώστε να μετατραπεί από string attribute σε nominal attribute. Στη διαδικασία που ακολουθείται στα πειράματα, η μετατροπή αυτή δεν συνεισφέρει κάπου καθώς οι πιθανές επιλογές που δίνονται στον χρήστη για class attributes είναι μόνο nominal attributes.

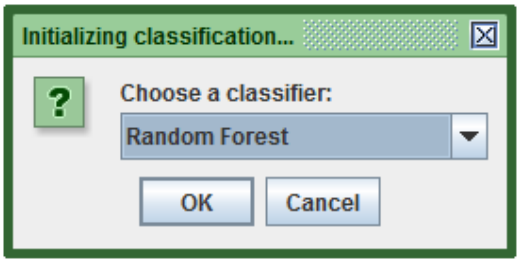
Το φίλτρο *StringToWordVector* εφαρμόζεται στη συνέχεια ταυτόχρονα και για τα δύο datasets σε μορφή παρτίδων. Το φίλτρο μετατρέπει string attributes σε ένα σύνολο από attributes που αναπαριστούν την συχνότητα της εμφάνισης των λέξεων από ένα κείμενο που περιλαμβάνεται στα string attributes. Για κάθε ιδιότητα τύπου δεδομένων που είναι τύπου string, εφαρμόζεται το φίλτρο και τα string attributes αντικαθίσταται από 1000 νέα attribute που η ονομασία τους ξεκινά με το όνομα του attribute όπως προϋπήρχε, ακολουθούμενη από την νέα λέξη που διατηρήθηκε. Κάθε λέξη μετατρέπεται πρώτα σε πεζά, για κάθε λέξη η έξοδος είναι η συχνότητα της λέξης παρά η απλή δυαδική δήλωση του αν παρίσταται. Ο *NullStemmer* έχει δηλωθεί ως αλγόριθμος παύσης (stammering algorithm). Ένα αρχείο που διατηρεί λέξεις προς εξαίρεση (stopwords) παρέχεται στο πρόγραμμα πριν την επιλογή του *NGramTokenizer* ως αλγορίθμου περικοπής λέξεων (tokenizing algorithm). Τα πιο κοινά διαχωριστικά ορίστηκαν μαζί με κάποια που προέκυψαν κατά τον έλεγχο της

απόδοσης του φίλτρου. Το φίλτρο δέχεται ως είσοδο το training dataset για να καθορίσει το σύνολο των λέξεων που θα δοθούν στην έξοδο και ύστερα εφαρμόζεται ταυτόχρονα στο training και στο testing dataset ώστε να έχει την ίδια επίδοση και στα δύο.

Μετά την προεπεξεργασία των dataset, αποθηκεύονται ως νέα ARFF files με τις ονομασίες *newTrainData.arff* και *newTestData.arff* αντίστοιχα.

4.3. Ταξινόμηση

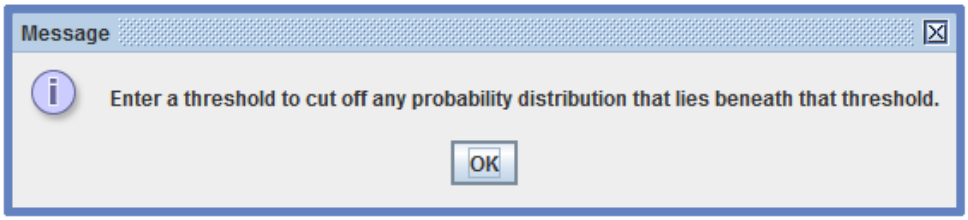
Μετά την προεπεξεργασία των datasets, εμφανίζεται το επόμενο παράθυρο που επιτρέπει στον χρήστη να επιλέξει ταξινομητή ανάμεσα στους *Random Forest* και *k-Nearest Neighbor*, όπως προαναφέρθηκε στο Κεφάλαιο 2.



Εικόνα 6 Ζητείται από τον χρήστη να επιλέξει classifier.

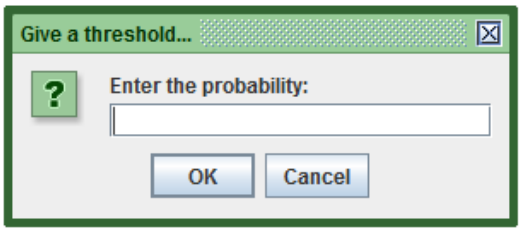
Ανάλογα με την επιλογή του χρήστη, ένας *Filtered Classifier* χρησιμοποιείται για να εκτελέσει τον επιλεγμένο ταξινομητή στα δεδομένα αφού έχουν φιλτραριστεί από το *RemoveType* Unsupervised Filter για την αφαίρεση των string attributes. Ο λόγος που πρώτα εφαρμόζεται το φίλτρο *RemoveType* τόσο στο training όσο και στο testing dataset είναι για την εξάλειψη του attribute των Individuals – το μοναδικό string attribute ύστερα από την εφαρμογή του φίλτρου *StringToWordVector* σε όλες τις ιδιότητες τύπου δεδομένων που ήταν τύπου string, και τις μετέτρεψε σε numeric attributes – έτσι ώστε να επιβεβαιωθεί πως οι ταξινομητές δεν θα δημιουργήσουν κατά λάθος κανόνες που θα εξαρτώνται από τα ονόματα των ατόμων.

Ο επιλεγμένος ταξινομητής οικοδομείται (built) βασιζόμενος στο training dataset και στη συνέχεια αξιολογείται με βάση το testing dataset. Για την λήψη της απόφασης σχετικά με το αν πρέπει να διατηρηθούν οι προβλέψεις για το class attribute ενός ταξινομημένου στιγμιότυπου, αξιολογείται η κατανομή της πιθανότητας κάθε πρόβλεψης. Ο χρήστης ενημερώνεται πως πρέπει να εισάγει ενός όριο (threshold) για να αποκλειστεί κάθε κατανομή πιθανότητας που βρίσκεται κάτω από το δοσμένο όριο.



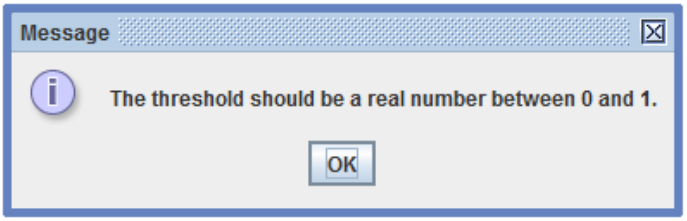
Εικόνα 7 Ο χρήστης ενημερώνεται πως στο επόμενο παράθυρο του ζητείται να εισάγει ένα threshold.

Στη συνέχεια, εμφανίζεται το παράθυρο για την εισαγωγή του ορίου. Δεδομένου ότι το όριο αφορά κατανομή πιθανότητας, ο χρήστης οφείλει ναι εισάγει ένα πραγματικό αριθμό στο διάστημα μηδέν έως ένα.



Εικόνα 8 Ζητείται από τον χρήστη να εισάγει ένα όριο κατανομής πιθανότητας.

Εάν το όριο δεν βρίσκεται στο επιθυμητό διάστημα τιμών, το επόμενο παράθυρο εμφανίζεται για να ενημερώσει τον χρήστη για το λάθος του, και στην πορεία επανεμφανίζεται το παράθυρο της Εικόνας 8 για την εισαγωγή του ορίου εκ νέου.



Εικόνα 9 Ο χρήστης πληροφορείται πως το όριο πρέπει να βρίσκεται στο διάστημα ανάμεσα σε 0 και 1.

Εν τέλει, οι τιμές που προέκυψαν από την πρόβλεψη του ταξινομητή και είχαν μεγαλύτερη κατανομή πιθανότητας από αυτή που εισήγαγε ο χρήστης, αποθηκεύονται στο training dataset και το νέο επισημασμένο (labeled) dataset αποθηκεύεται στο αρχείο με ονομασία *labeled.arff*.

4.4. Εμπλουτισμός της Οντολογίας με Ταξινόμηση

Μετά την εφαρμογή της ταξινόμησης ανάλογα με το class attribute και τον αλγόριθμο ταξινόμησης που επέλεξε ο χρήστης, η αρχική οντολογία μπορεί να εμπλουτιστεί με τα αποτελέσματα της ταξινόμησης. Αυτή η διαδικασία εμπλουτίζει το ABox της οντολογίας.

Κατ' αρχάς, διαφορετικοί ισχυρισμοί θα εμπλουτίσουν την οντολογία ανάλογα με το κατά πόσο το class attribute είναι αναπαράσταση κλάσης ή ιδιότητας αντικειμένου της οντολογίας. Στη συνέχεια, οι ισχυρισμοί εξαρτώνται από την τιμή της πρόβλεψης για το class attribute του κάθε στιγμιότυπου. Εάν η πρόβλεψη δεν ήταν αρκετά καλή, με κατανομή πιθανότητας μικρότερη από το όριο που εισήγαγε ο χρήστης, τότε θα τοποθετηθεί μία τιμή Missing Value αντί κάποιας δυαδικής τιμής και δεν υπάρχει ισχυρισμός που μπορεί να εμπλουτίσει την οντολογία. Όμως εάν η τιμή του class attribute του στιγμιότυπου είναι "**true**" ή "**false**", τότε από αυτή την τιμή θα εξαρτηθεί τι είδους ισχυρισμός θα προστεθεί στην οντολογία.

Εάν το class attribute αντιπροσωπεύει μία κλάση της οντολογίας, τότε ο μόνος δυνατός ισχυρισμός είναι όταν η τιμή του πρόσφατα επισημασμένου στιγμιότυπου είναι "**true**" καθώς δεν υπάρχει αρνητικός ισχυρισμός κλάσης. Επομένως, έστω ένα στιγμιότυπο με την τιμή $individual$ να αντιστοιχεί στο attribute των *Individuals*, που έχει το class attribute $class$ επισημασμένο ως "**true**", τότε ο ακόλουθος ισχυρισμός εισάγεται στην οντολογία:

$$OWLClassAssertionAxiom(class, individual)$$

Όταν το class attribute αντιπροσωπεύει είτε ιδιότητα αντικειμένου είτε αντίστροφη ιδιότητα αντικειμένου υπάρχουν διαφορετικές περιπτώσεις ανάλογα με την επιλογή του χρήστη να διατηρήσει τις σχέσεις ως υπαρξιακές ή όχι.

Στην περίπτωση που οι σχέσεις έχουν διατηρηθεί ως υπαρξιακές και εάν το class attribute αντιπροσωπεύει μια ιδιότητα αντικειμένου της οντολογίας, τότε οι πιθανοί ισχυρισμοί υφίστανται όταν η τιμή του πρόσφατα επισημασμένου στιγμιότυπου είναι είτε "**true**" είτε "**false**". Επομένως, έστω ένα στιγμιότυπο με την τιμή ind να αντιστοιχεί το attribute των *Individuals*, που έχει το class attribute $objProp$ επισημασμένο ως "**true**", τότε ο ακόλουθος ισχυρισμός εισάγεται στην οντολογία:

$$OWLObjectPropertyAssertionAxiom(objProp, ind, ind2)$$

όπου η τιμή $ind2$ αντιστοιχεί στο $OWLAnonymousIndividual$.

Έστω ένα στιγμιότυπο με την τιμή ind να αντιστοιχεί το attribute των *Individuals*, που έχει το class attribute $objProp$ επισημασμένο ως "**false**", τότε ο ακόλουθος ισχυρισμός εισάγεται στην οντολογία:

$$OWLNegativeObjectPropertyAssertionAxiom(objProp, ind, ind2)$$

όπου η τιμή $ind2$ αντιστοιχεί στο $OWLAnonymousIndividual$.

Εάν το class attribute αντιπροσωπεύει μια αντίστροφη ιδιότητα αντικειμένου της οντολογίας, τότε οι πιθανοί ισχυρισμοί υφίστανται όταν η τιμή του πρόσφατα επισημασμένου στιγμιότυπου είναι είτε "**true**" είτε "**false**". Επομένως, έστω ένα στιγμιότυπο με την τιμή ind να αντιστοιχεί το attribute των *Individuals*, που έχει το class attribute $invObjProp$ επισημασμένο ως "**true**", τότε ο ακόλουθος ισχυρισμός εισάγεται στην οντολογία:

$$OWLObjectPropertyAssertionAxiom(invObjProp, ind2, ind)$$

όπου η τιμή $ind2$ αντιστοιχεί στο $OWLAnonymousIndividual$.

Έστω ένα στιγμιότυπο με την τιμή ind να αντιστοιχεί το attribute των *Individuals*, που έχει το class attribute $invObjProp$ επισημασμένο ως "**false**", τότε ο ακόλουθος ισχυρισμός εισάγεται στην οντολογία:

$$OWLNegativeObjectPropertyAssertionAxiom(invObjProp, ind2, ind)$$

όπου η τιμή $ind2$ αντιστοιχεί στο $OWLAnonymousIndividual$.

Στην περίπτωση που οι σχέσεις δεν έχουν διατηρηθεί ως υπαρξιακές και εάν το class attribute αντιπροσωπεύει μια ιδιότητα αντικειμένου της οντολογίας, τότε οι πιθανοί ισχυρισμοί υφίστανται όταν η τιμή του πρόσφατα επισημασμένου στιγμιότυπου είναι είτε "**true**" είτε "**false**". Επομένως, έστω ένα στιγμιότυπο με την τιμή ind να αντιστοιχεί το attribute των *Individuals*, που έχει το class attribute $objProp$ επισημασμένο ως "**true**", τότε ο ακόλουθος ισχυρισμός εισάγεται στην οντολογία:

$$OWLObjectPropertyAssertionAxiom(objProp, ind, ind2)$$

όπου η τιμή του $ind2$ αντιστοιχεί στο όνομα του ατόμου που έχει διατηρηθεί στο όνομα του class attribute.

Έστω ένα στιγμιότυπο με την τιμή ind να αντιστοιχεί το attribute των *Individuals*, που έχει το class attribute $objProp$ επισημασμένο ως "**false**", τότε ο ακόλουθος ισχυρισμός εισάγεται στην οντολογία:

$$OWLNegativeObjectPropertyAssertionAxiom(objProp, ind, ind2)$$

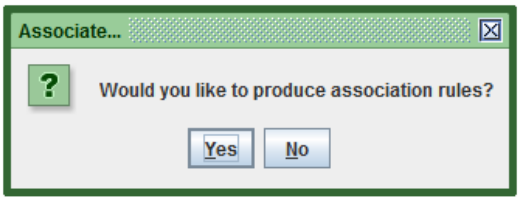
όπου η τιμή του $ind2$ αντιστοιχεί στο όνομα του ατόμου που έχει διατηρηθεί στο όνομα του class attribute.

Εν τέλει, οι ισχυρισμοί που προέκυψαν εισάγονται στην αρχική οντολογία, και η νέα εμπλουτισμένη οντολογία αποθηκεύεται με την ονομασία *Enriched.owl*.

4.5. Συσχέτιση

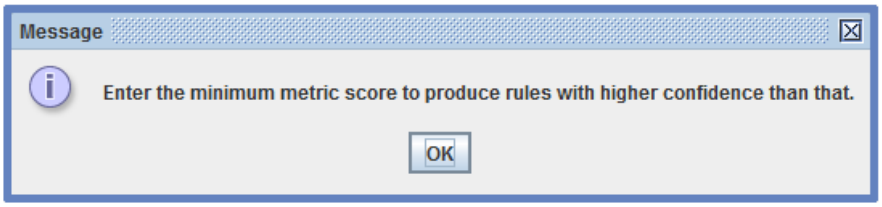
Η ιδέα του εμπλουτισμού μιας οντολογίας χωρίς την χρήση εξωτερική βοήθειας από διαφορετικές οντολογίες να χρησιμεύουν ως training datasets επιτυγχάνεται με τη εκμάθηση κανόνων συσχέτισης. Αυτοί οι αλγόριθμοι παράγουν κανόνες βασιζόμενοι μόνο στα attributes που αποτελούν στιγμιότυπα ενός dataset, αυτό που είχε προηγουμένως αναφερθεί και ως testing dataset.

Κατά την εκτέλεση του προγράμματος και αφού έχει εισαχθεί η οντολογία που χρησιμοποιείται αργότερα ως testing dataset, ο χρήστης ερωτάται για την παραγωγή κανόνων συσχέτισης από το παράθυρο που παρουσιάζεται στην Εικόνα 10.



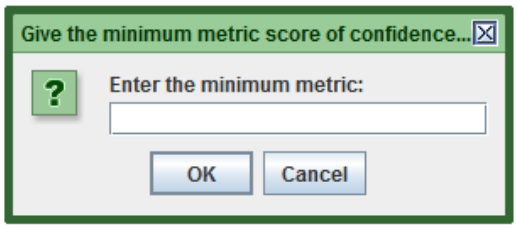
Εικόνα 10 Ο χρήστης ερωτάται για την παραγωγή association rules.

Στη συνέχεια, ζητείται από τον χρήστη να εισάγει ένα όριο για το βαθμό της μετρικής (metric score). Στο παρόν μοντέλο, ο τύπος της μετρικής που επιλέχθηκε είναι η *Confidence*, και επομένως ζητείται από τον χρήστη να εισάγει ένα ελάχιστο metric score για την εξάλειψη κανόνων που πιθανόν να έχουν μικρότερη *Confidence*.



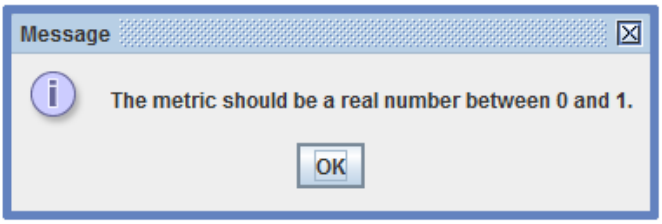
Εικόνα 11 Ο χρήστης πληροφορείται πως στο επόμενο παράθυρο πρέπει να εισάγει το minimum metric score.

Ακολούθως, το παράθυρο για την εισαγωγή του ελάχιστου metric score εμφανίζεται. Καθώς το όριο αυτό αφορά την μετρική *Confidence* metric ο χρήστης πρέπει να εισάγει έναν πραγματικό αριθμό στο διάστημα ανάμεσα σε μηδέν και ένα.



Εικόνα 12 Ζητείται από τον χρήστη να εισάγει το minimum metric score.

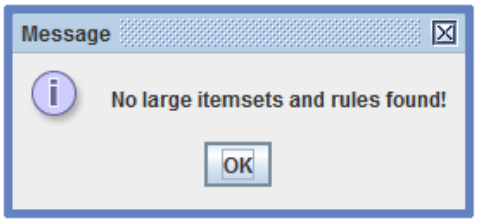
Εάν το όριο δεν βρίσκεται στο διάστημα ανάμεσα στο μηδέν και το ένα, εμφανίζεται το επόμενο παράθυρο για να πληροφορήσει το χρήστη πως η τιμή που εισήγαγε είναι λάθος, και ακολούθως επανεμφανίζεται το παράθυρο που απεικονίζεται στην Εικόνα 12 για την εισαγωγή νέου ορίου.



Εικόνα 13 Ο χρήστης ενημερώνεται πως το minimum metric score πρέπει να βρίσκεται στο διάστημα ανάμεσα σε 0 και 1.

Στο παρόν μοντέλο, ο αλγόριθμος συσχέτισης που χρησιμοποιείται για τον εμπλουτισμό του TBox της αρχικής οντολογίας είναι ο *Apriori*, όπως προαναφέρθηκε στο Κεφάλαιο 2. Το ελάχιστο metric score έχει δοθεί από τον χρήστη. Ο αλγόριθμος *Apriori* μπορεί να λειτουργήσει μόνο με nominal attributes, επομένως για την παραγωγή κανόνων συσχέτισης και την αποφυγή τυχών προβλημάτων, εφαρμόζεται το φίλτρο *RemoveType* ώστε να απαλείψει attributes τύπου string και numeric τα οποία πιθανόν να συνθέτουν το dataset εκτός από τα τύπου nominal.

Εν τέλει, οι συσχετισμοί οικοδομούνται και παράγονται κανόνες οι οποίοι δίνονται ως παράμετρος στο τελικό κομμάτι του προγράμματος που εμπλουτίζει το TBox της οντολογίας. Στην σπάνια περίπτωση που δεν παράγεται κανένας κανόνας επειδή δεν υπάρχουν μεγάλα σύνολα στοιχείων (itemsets) στο dataset, ο χρήστης ενημερώνεται από το παράθυρο που ακολουθεί.



Εικόνα 14 Ο χρήστης ενημερώνεται πως δεν υπήρξαν μεγάλα itemsets για την παραγωγή κανόνων.

4.6. Εμπλουτισμός της Οντολογίας με Κανόνες Συσχέτισης

Μετά την απόκτηση των κανόνων συσχέτισης, η αρχική οντολογία μπορεί να εμπλουτιστεί με ισχυρισμούς που βασίζονται στο αντικείμενο που αντιπροσωπεύεται από κάθε attribute που συμμετέχει στον κανόνα που προέκυψε. Αυτή η διαδικασία εμπλουτίζει το TBox της οντολογίας.

Διάφοροι τύπου κανόνων ελέγχθηκαν λεπτομερώς και αυτοί που είχαν υψηλή *Confidence* και ήταν δυνατό να χειριστούν, εισάχθηκαν στην οντολογία με τη μορφή ισχυρισμών. Οι ισχυρισμοί εξαρτώνται από την ποσότητα των attributes που απαρτίζουν κάθε κανόνα, την τιμή αυτών των attributes και τα αντικείμενα της οντολογίας στα οποία αντιστοιχούν. Επιτεύχθηκε ο χειρισμός μόνο των κανόνων που είχαν ένα attribute στο πρώτο μέρος τους (if … then …). Οι ακόλουθοι τύπου κανόνων παρατηρήθηκαν:

- If $attribute1$ = **true**, then $attribute2$ = **true**

Σε αυτή την περίπτωση οι ισχυρισμοί κλάσης που προκύπτουν από τα δύο attributes, $clExpression1$ και $clExpression2$ αντίστοιχα, εισάχθηκαν στην οντολογία με την μορφή του ακόλουθου ισχυρισμού:

$$OWLSubClassOfAxiom(clExpression1, clExpression2)$$

- If $attribute1$ = **true**, then $attribute2$ = **false**

Σε αυτή την περίπτωση οι ισχυρισμοί κλάσης που προκύπτουν από τα δύο attributes, $clExpression1$ και $clExpression2$ αντίστοιχα, εισάχθηκαν στην οντολογία με την μορφή του ακόλουθου ισχυρισμού:

$$OWLDisjointClassesAxiom(clExpression1, clExpression2)$$

- If $attribute1$ = **false**, then $attribute2$ = **true**

Σε αυτή την περίπτωση οι ισχυρισμοί κλάσης που προκύπτουν από τα δύο attributes, $clExpression1$ και $clExpression2$ αντίστοιχα, εισάχθηκαν στην οντολογία με την μορφή του ακόλουθου ισχυρισμού:

$$OWLSubClassOfAxiom\big(owl{:}Thing, UnionOf\,(clExpression1, clExpression2)\big)$$

- If $attribute1$ = **false**, then $attribute2$ = **false**

Σε αυτή την περίπτωση οι ισχυρισμοί κλάσης που προκύπτουν από τα δύο attributes, $clExpression1$ και $clExpression2$ αντίστοιχα, εισάχθηκαν στην οντολογία με την μορφή του ακόλουθου ισχυρισμού:

$$OWLSubClassOfAxiom(clExpression2, clExpression1)$$

Κάθε ισχυρισμός κλάσης (class expression) προέκυψε όπως περιγράφεται ακολούθως, εξαρτώμενος από τα attributes και τα αντικείμενα που αντιστοιχούν:

Εάν το attribute αντιστοιχεί σε μια κλάση, τότε ο ισχυρισμός κλάσης είναι η $OWLClass$ που αναφέρεται στο όνομα του attribute.

Εάν το attribute αντιστοιχεί σε μια ιδιότητα αντικειμένων όπου οι σχέσεις έχουν διατηρηθεί ως υπαρξιακές, τότε ο ισχυρισμός κλάσης είναι ο ακόλουθος:

$$OWLObjectSomeValuesFrom(objProperty, owl{:}Thing)$$

όπου $objProperty$ είναι η $OWLObjectProperty$ που αναφέρεται στο όνομα του attribute.

Εάν το attribute αντιστοιχεί σε μια ιδιότητα αντικειμένων όπου οι σχέσεις δεν έχουν διατηρηθεί ως υπαρξιακές, τότε ο ισχυρισμός κλάσης είναι ο ακόλουθος:

$$OWLObjectHasValue(objProperty, individual)$$

όπου $objProperty$ είναι η $OWLObjectProperty$ που αναφέρεται στο όνομα του attribute και $individual$ είναι το όνομα του ατόμου που έχει εξίσου διατηρηθεί στο όνομα του attribute.

Εάν το attribute αντιστοιχεί σε μια αντίστροφη ιδιότητα, τότε ο ισχυρισμός κλάσης είναι ο ακόλουθος:

$$OWLObjectSomeValuesFrom(invObjProperty, owl{:}Thing)$$

όπου $invObjProperty$ είναι η $OWLObjectInverseOf(objProperty)$ και $objProperty$ είναι η $OWLObjectProperty$ που αναφέρεται στο όνομα του attribute.

Εν τέλει, οι ισχυρισμοί που προέκυψαν προστέθηκαν στην αρχική οντολογία, και η νέα εμπλουτισμένη οντολογία αποθηκεύεται με το όνομα *Enriched2.owl*.

Κεφάλαιο 5. Πειραματικά Αποτελέσματα

Σε αυτό το κεφάλαιο, περιγράφονται τα δεδομένα που χρησιμοποιήθηκαν σαν είσοδος στο σύστημα που παρουσιάστηκε στο Κεφάλαιο 4, καθώς και η έξοδος του συστήματος μετά την διεξαγωγή των πειραμάτων.

5.1. Οντολογίες Εισόδου

Το σύστημα που παρουσιάστηκε προηγουμένως εκτελείται φορτώνοντας αρχικά δύο οντολογίες που θα χρησιμοποιηθούν ως training dataset και μία οντολογία ως testing dataset, η οποία είναι αυτή που θα εμπλουτιστεί.

Οι οντολογίες έχουν δημιουργηθεί από συλλογές που είναι τμήματα της πλατφόρμας WITH [17] για πολιτισμικά δεδομένα, που εκθέτει APIs από διαφορετικές πύλες και αποθήκες (portals and repositories). Οι δύο πρώτες οντολογίες αντιστοιχούν στις συλλογές “F&D TopFoto” (Food & Drink) [18] και “TopFoto EuropeanaPhotography”

[19] από τη Europeana Collections [20]. Η οντολογία που θα εμπλουτιστεί αντιστοιχεί στη συλλογή “Related to Food and Drink” [21] από τη Europeana Collections.

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του ταξινομητή προέρχονται από την TopFoto που είναι μια ανεξάρτητη βιβλιοθήκη εικόνων με έδρα το Edenbridge που βρίσκεται 45 λεπτά νότια του Λονδίνου. Συνολικά το αρχείο τους περιλαμβάνει πάνω από 10 εκατομμύρια εικόνες από μεσαιωνικά έγγραφα έως σημερινή ψηφιακά αρχεία που αποστέλλονται μέσω FTP στους πελάτες τους σε ολόκληρο τον κόσμο. Ο πυρήνας του αρχείου βρίσκεται σε έντυπη μορφή αρχείου και αποτελείται από 120.000 αρνητικά από τον John Topham (που είναι φωτογράφος και ιδρυτής του TopFoto) καθώς και από εκατομμύρια αρνητικών και έντυπης μορφής εκτυπώσεων από διάφορα ιστορικά πρακτορεία τύπου που έχουν συλλεχθεί από το 1975 από τον σημερινό ιδιοκτήτη, Alan Smith (πρόεδρος της CEPIC 1997-2009). Το αρχείο του TopFoto τροφοδοτεί κυρίως πελάτες που δραστηριοποιούνται στον έντυπο τύπο, αλλά είναι εξαιρετικά ποικιλόμορφο και οι εικόνες του έχουν επιτυχημένα χρησιμοποιηθεί σε όλους τους τομείς της οπτικής δημοσίευσης (εφημερίδες, περιοδικά, αφίσες, καρτποστάλ κα).

Το TopFoto ήταν πρωτοπόρος στην ψηφιοποίηση και την ηλεκτρονική μεταφορά περιεχομένου μέσω των νέων τεχνολογιών που έχουν προκύψει και έχει στενούς δεσμούς με διεθνείς εταίρους σε περισσότερες από 40 χώρες σε όλο τον κόσμο.

Πιο συγκεκριμένα για την εργασία μας χρησιμοποιήθηκε το σύνολο του περιεχομένου που η βιβλιοθήκη TopFoto δημοσίευσε στη Europeana για δύο θεματικά έργα το “EuropeanaPhotography” και το “Europeana Food and Drink”.

Ο κύριος σκοπός του “EuropeanaPhotography” ήταν να προετοιμάσει και να εξασφαλίσει την ποιότητα περιεχομένου καθώς και να συνεισφέρει πάνω από 430.000 φωτογραφικά αντικείμενα στην Europeana, που αντιπροσωπεύουν από κοινού μια επιλογή από αριστουργήματα από την αρχή της φωτογραφικής ιστορίας. Έτσι λοιπόν το περιεχόμενο του TopFoto που δημοσιεύτηκε για αυτό το έργο και χρησιμοποιήθηκε για την δημιουργία του ταξινομητή αποτελείται από 60,881 εικόνες που θεματικά σχετίζονται με χώρους (πόλεις - όπως η μετατροπή του Παρισιού από τον Haussmann και της Βαρκελώνης από τον Gaudi, τοπία - όπως το ευρωπαϊκό τοπίο το 1800 κλπ ...), άτομα (πορτρέτα - Queen Victoria, οι πάπες, Garibaldi, Coco Chanel - και την καθημερινή ζωή), γεγονότα (πολιτικά γεγονότα - la Commune de Paris εμφύλιοι πόλεμοι, βασιλικοί γάμοι, κ.λπ.) και «τάσεις» ή «κινήσεις» (βιομηχανική επανάσταση, χειραφέτηση, καλλιτεχνικά ρεύματα, γεωγραφικές εξερευνήσεις, αποικισμός, κλπ).

Από την άλλη το περιεχόμενο του TopFoto για το “Europeana Food and Drink”, όπως ορίζει και το όνομα του έργου, αποτελείται από 7,837 αρχεία που σχετίζονται με την γαστρονομική κληρονομιά της Ευρώπης. Τέλος, η τρίτη συλλογή που χρησιμοποιήθηκε για να αξιολογηθεί ο ταξινομητής που κατασκευάσαμε δημιουργήθηκε κάνοντας χρήση ενός θησαυρού (ιεραρχία εννοιών) που κατασκευάστηκε από το έργο και το API της Europeana. Πιο συγκεκριμένα ένα

υποσύνολο των εννοιών του θησαυρού χρησιμοποιήθηκαν ως όροι αναζήτησης στη Europeana, μαζί με το κριτήριο το έργο από το οποίο προήλθαν να μην είναι το “Europeana Food and Drink”. Με αυτό τον τρόπο το σύνολο του περιεχομένου που συσσωρεύτηκε στην “Related to Food and Drink” συλλογή χαρακτηρίστηκε σαν σχετικό με την γαστρονομική κληρονομιά της Ευρώπης αλλά χωρίς να προέρχεται από το συγκεκριμένο έργο και σκοπός μας με την χρήση του ταξινομητή ήταν να ξεκαθαρίσουμε την συλλογή αυτή από μη σχετικό με την γαστρονομική κληρονομιά της Ευρώπης περιεχόμενο.

Επομένως, ο πρακτικός στόχος της ταξινόμησης σε αυτό το μοντέλο είναι να προβλέψουμε κατά πόσο τα στοιχεία της συλλογής “Related to Food and Drink” ανήκουν στο “Food and Drink Project” γνωρίζοντας ότι η συλλογή “F&D TopFoto” ανήκει σε αυτό το Project και επίσης έχοντας ως δεδομένο ότι η συλλογή “TopFoto EuropeanaPhotography” δεν έχει στοιχεία που ανήκουν στο “Food and Drink Project”. Η έξοδος του συστήματος που παρουσιάστηκε είναι μία νέα, εμπλουτισμένη οντολογία που δημιούργησε μία νέα κλάση που αντιστοιχεί στη συλλογή “F&D TopFoto” για τα στοιχεία τα οποία δεν ήταν μέρος της προαναφερθείσας συλλογής ενώ θα έπρεπε να είναι.

Επιπροσθέτως, παράχθηκαν κανόνες συσχέτισης από την οντολογία που αντιστοιχεί στη συλλογή “Related to Food and Drink”. Οι κανόνες αυτοί επισημαίνουν συσχετίσεις μεταξύ κλάσεων και ιδιοτήτων αντικειμένων της οντολογίας, οι οποίες νέες συσχετίσεις προστίθενται αργότερα για να εμπλουτίσουν περαιτέρω την αρχική οντολογία.

5.2. Αξιολόγηση της Ταξινόμησης

Το αρχείο ARFF που δημιουργήθηκε από τις δύο οντολογίες που αντιστοιχούν στις συλλογές “F&D TopFoto” και “TopFoto EuropeanaPhotography” χρησιμοποιήθηκε σαν training dataset για τον ταξινομητή που ο χρήστης επέλεξε να εφαρμόσει. Για να αξιολογηθεί η απόδοση του κάθε ταξινομητή, εφαρμόστηκε ταξινόμηση σε ποσοστιαίο διαχωρισμό του συνόλου δεδομένων εκπαίδευσης (training dataset) και τα αποτελέσματα για τον κάθε ταξινομητή παρουσιάζονται στη συνέχεια.

Πρώτα από όλα, αξιολογείται η απόδοση του ταξινομητή RandomForest, όπου ο τρόπος ελέγχου αποτελείται από ποσοστιαίο διαχωρισμό κατά τον οποίο ένα ποσοστό του συνόλου δεδομένου χρησιμοποιείται για εκπαίδευση και το υπόλοιπο χρησιμοποιείται για έλεγχο. Παρουσιάζεται παρακάτω η αξιολόγηση ανά ποσοστό διαχωρισμού και ο confusion matrix:

- Split 70% train, remainder test:

  === Evaluation on test split ===

Correctly Classified Instances	**3709**	**99.8116** %
Incorrectly Classified Instances	7	0.1884 %

  === Confusion Matrix ===

Classified as	**a**	**b**
a = true	2361	6
b = false	1	1348

- Split 50% train, remainder test:

  === Evaluation on test split ===

Correctly Classified Instances	**6173**	**99.661** %
Incorrectly Classified Instances	21	0.339 %

  === Confusion Matrix ===

Classified as	**a**	**b**
a = true	3902	18
b = false	3	2271

- Split 30% train, remainder test:

  === Evaluation on test split ===

Correctly Classified Instances	**8624**	**99.4465** %
Incorrectly Classified Instances	48	0.5535 %

=== Confusion Matrix ===

Classified as	a	b
a = true	5434	42
b = false	6	3190

Επιπλέον, η απόδοση του ταξινομητή kNN αξιολογείται με τον ίδιο ακριβώς τρόπο με προηγουμένως. Ακολουθεί η αξιολόγηση ανά ποσοστό διαχωρισμού και ο confusion matrix:

- Split 70% train, remainder test:

=== Evaluation on test split ===

Correctly Classified Instances	**3666**	**98.6545** %
Incorrectly Classified Instances	50	1.3455 %

=== Confusion Matrix ===

Classified as	a	b
a = true	2324	43
b = false	7	1342

- Split 50% train, remainder test:

=== Evaluation on test split ===

Correctly Classified Instances	**6085**	**98.2402** %
Incorrectly Classified Instances	109	1.7598 %

=== Confusion Matrix ===

Classified as	a	b
a = true	3824	96
b = false	13	2261

- Split 30% train, remainder test:

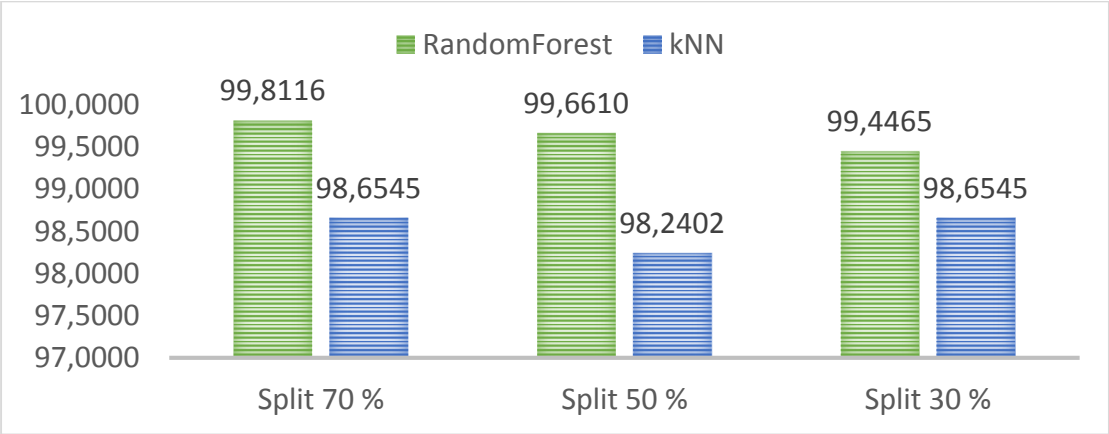
  === Evaluation on test split ===

Correctly Classified Instances	**3666**	**98.6545 %**
Incorrectly Classified Instances	50	1.3455 %

=== Confusion Matrix ===

Classified as	**a**	**b**
a = true	2324	43
b = false	7	1342

Από την ακρίβεια είναι προφανές ότι οι ταξινομητές πιθανόν να προσαρμόζονται υπερβολικά στα δεδομένα (overfitting) καθώς ακόμη και με πολύ μικρό ποσοστό των δεδομένων για εκπαίδευση η ακρίβεια αγγίζει το 100%. Συγκεντρωτικά, η ακρίβεια του κάθε ταξινομητή ανά ποσοστό διαχωρισμού δίνεται στο ακόλουθο διάγραμμα:



Το επόμενο βήμα είναι να αξιολογήσουμε τα attributes σε σχέση με το class attribute. Για την επίτευξη αυτού χρησιμοποιείται ο InfoGain Attribute Selector ώστε να αξιολογήσει την αξία των attributes μετρώντας το κέρδος πληροφορίας σε σχέση με την κλάση, όπου η μέθοδος αναζήτησης είναι η Ranker, η οποία κατατάσσει τα χαρακτηριστικά με βάση τις ατομικές τους αξιολογήσεις. Το αποτέλεσμα της αξιολόγησης για τα πρώτα χαρακτηριστικά που έδειξαν κέρδος άνω του 0.1 είναι τα ακόλουθα:

InfoGain	Attribute
0.948633	hasContributor_objectProperty
0.744553	hasKeywords_dataProperty_food
0.729629	hasKeywords_dataProperty_drink
0.689258	hasKeywords_dataProperty_photography
0.670505	hasKeywords_dataProperty_monochrome
0.641136	hasKeywords_dataProperty_century
0.636695	hasKeywords_dataProperty_black
0.549799	hasKeywords_dataProperty_white
0.495657	hasKeywords_dataProperty_wars
0.489837	hasKeywords_dataProperty_eufd
0.487827	hasKeywords_dataProperty_twentieth
0.450403	hasDescription_dataProperty_topfoto
0.450403	hasLabel_dataProperty_topfoto
0.446517	hasLabel_dataProperty_credit
0.446517	hasDescription_dataProperty_credit
0.446001	hasDescription_dataProperty_thepicturekitchen
0.446001	hasLabel_dataProperty_thepicturekitchen
0.380787	hasKeywords_dataProperty_nineteen
0.354535	hasKeywords_dataProperty_th
0.344098	hasDescription_dataProperty_avery
0.344098	hasLabel_dataProperty_avery
0.342283	hasLabel_dataProperty_louise
0.342283	hasDescription_dataProperty_louise
0.34015	hasLabel_dataProperty_marie
0.34015	hasDescription_dataProperty_marie
0.277721	hasKeywords_dataProperty_thirties
0.211142	hasKeywords_dataProperty_alfieri
0.105334	hasKeywords_dataProperty_foo

Επίσης κάποια τυχαία χαρακτηριστικά που έδειξαν μηδενικό κέρδος είναι:

InfoGain	Attribute
0	hasLabel_dataProperty_rooms
0	hasDescription_dataProperty_man
0	hasLabel_dataProperty_ropes
0	hasDescription_dataProperty_make
0	hasDescription_dataProperty_mediterranean
0	hasLabel_dataProperty_run
0	hasDescription_dataProperty_various
0	hasKeywords_dataProperty_pest
0	hasDescription_dataProperty_mobile
0	hasDescription_dataProperty_usually
0	hasLabel_dataProperty_council

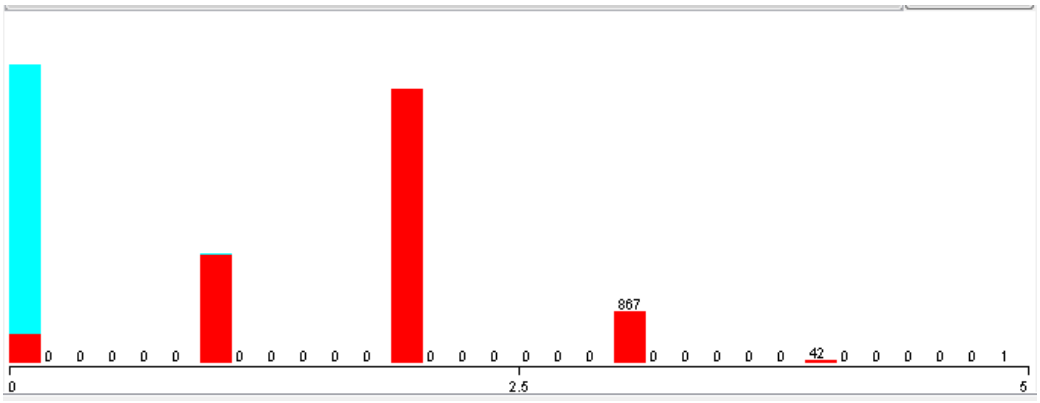
Από τα προαναφερθέντα καταταγμένα χαρακτηριστικά υπάρχουν κάποιες παρατηρήσεις:

Κατ' αρχάς, είναι εύκολο να παρατηρήσουμε ότι τα attributes με μηδενικό κέρδος δεν έχουν καμία σχέση με την επιθυμητή κλάση, καθώς καμία από τις λέξεις δεν δείχνει ομοιότητα με κάποια πιθανή συλλογή "Food & Drink".

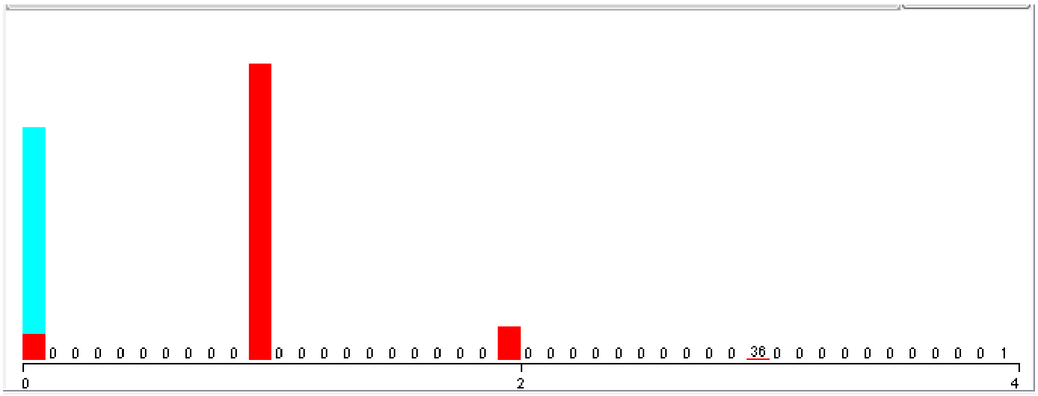
Δεύτερον, υπάρχουν παραπάνω από ένα attributes που είναι υψηλώς συσχετισμένα με το χαρακτηριστικό της κλάσης και τα οποία αντιστοιχούν στην ίδια λέξη. Παραδείγματος χάρη, τα hasDescription_dataProperty_topfoto και hasLabel_dataProperty_topfoto, τα οποία έχουν το ίδιο κέρδος καταδεικνύουν ότι οι ιδιότητες δεδομένων hasDescription και hasLabel έχουν σχέση με το "topFoto". Εμφανώς, μόνο ένα από τα δύο attributes θα ήταν αρκετό, αλλά καθώς και τα δύο ανήκουν στο dataset, ο ταξινομητής αποπροσανατολίζεται και η προκύπτουσα ταξινόμηση δεν είναι τόσο ακριβής.

Τέλος, το χαρακτηριστικό hasContributor_objectProperty έχει υψηλό κέρδος καθώς όλα τα στιγμιότυπα που ανήκουν στη συλλογή "F&D TopFoto" έχουν συγκεκριμένο συντελεστή (contributor), ενώ τα στιγμιότυπα της "TopFoto EuropeanaPhotography" δεν προσδιορίζουν κάποιον συντελεστή. Η παρουσία αυτού του attribute δείχνει υψηλή συσχέτιση με την επιθυμητή κλάση που λειτουργεί σαν διαταραχή στην ταξινόμηση που εκτελέστηκε. Ως αποτέλεσμα, αυτό το attribute διαγράφεται επίτηδες για να διευκολύνει την ταξινόμηση.

Η κατανομή για τα attributes που έχουν υψηλό κέρδος φαίνεται παρακάτω. Το πρώτο γράφημα δείχνει την κατανομή για το hasKeywords_dataProperty_food και το δεύτερο την κατανομή για το hasKeywords_dataProperty_drink.

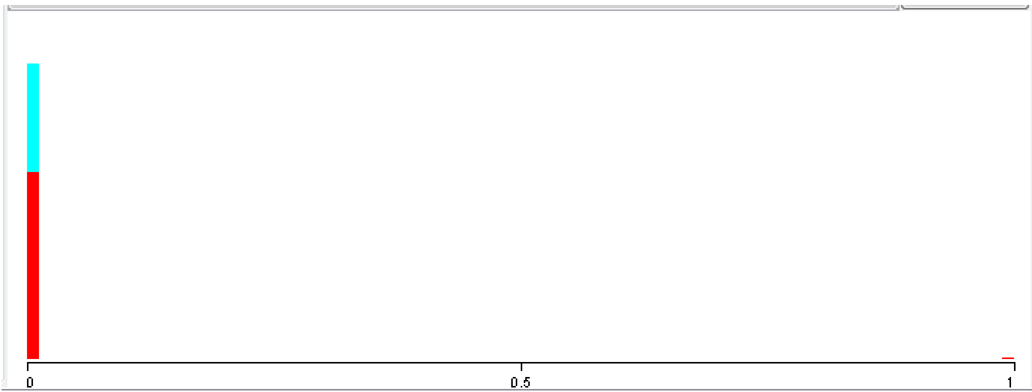


Εικόνα 15 Κατανομή του hasKeywords_dataProperty_food attribute με 0.744553 InfoGain (με κόκκινο φαίνονται τα στιγμιότυπα που ανήκουν στην επιθυμητή κλάση, με κυανοπράσινο αλλιώς)



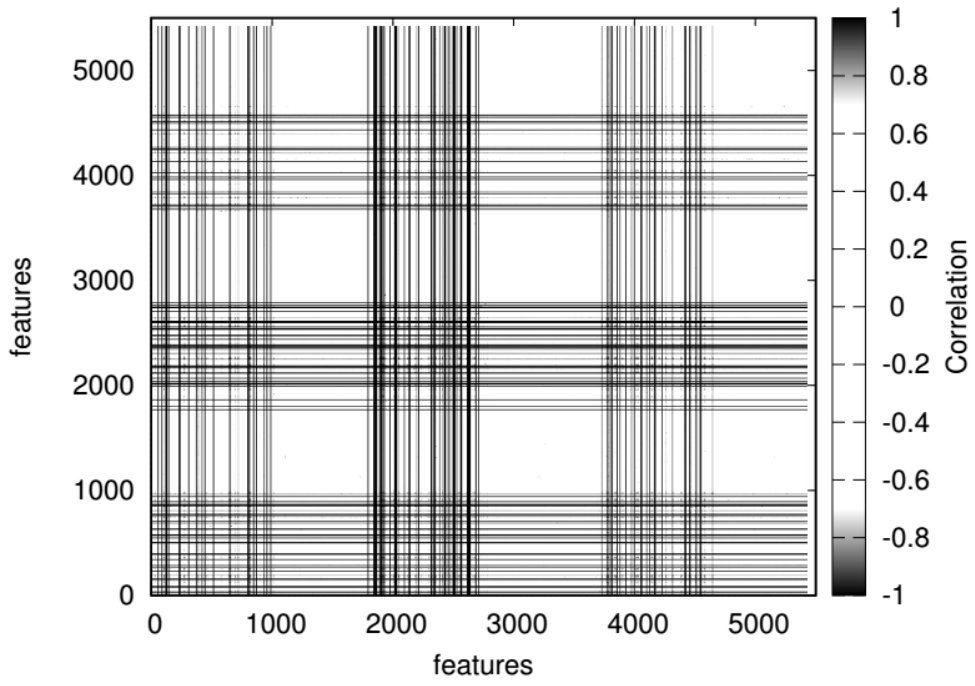
Εικόνα 16 Κατανομή τουhasKeywords_dataProperty_drink attribute με 0.729629 InfoGain (με κόκκινο φαίνονται τα στιγμιότυπα που ανήκουν στην επιθυμητή κλάση, με κυανοπράσινο αλλιώς)

Σε αντίθεση, η διαφορά στην κατανομή είναι εμφανής για ένα attribute με μηδενικό κέρδος όπως το hasDescription_dataProperty_modern.



Εικόνα 17 Κατανομή τουhasKeywords_dataProperty_modern attribute με 0 InfoGain (με κόκκινο φαίνονται τα στιγμιότυπα που ανήκουν στην επιθυμητή κλάση, με κυανοπράσινο αλλιώς)

Για την οπτικοποίηση της συσχέτισης μεταξύ των χαρακτηριστικών και για την ενίσχυση του ισχυρισμού πως είναι τόσο υψηλά συσχετισμένα που η ταξινόμηση δεν λειτουργεί όπως αναμενόταν, παρουσιάζεται παρακάτω ο Pearson Correlation Matrix.



Εικόνα 18 Pearson Correlation Matrix

Ο συντελεστής συσχέτισης Pearson λαμβάνεται διαιρώντας τη συνδιακύμανση των δύο μεταβλητών με το γινόμενο των τυπικών τους αποκλίσεων και ορίζεται μόνο αν και οι δύο τυπικές αποκλίσεις είναι διάφορες του μηδενός. Η συσχέτιση Pearson είναι +1 σε περίπτωση μίας τέλειας, απευθείας (αυξανόμενης) γραμμικής σχέσης (συσχέτιση – correlation), -1 σε περίπτωση μίας τέλειας, φθίνουσας (αντίστροφης) γραμμικής σχέσης (αντισυσχέτιση – anticorrelation), και κάποια τιμή μεταξύ -1 και 1 σε κάθε άλλη περίπτωση, καταδεικνύοντας το βαθμό γραμμικής εξάρτησης μεταξύ των μεταβλητών. Καθώς πλησιάζει το μηδέν υπάρχει όλο και λιγότερη σχέση (πιο κοντά στο να είναι ασυσχέτιστες). Όσο πιο κοντά είναι ο συντελεστής είτε στο -1, είτε στο 1, τόσο πιο δυνατή είναι η συσχέτιση μεταξύ των μεταβλητών. Στον πίνακα που παρουσιάστηκε, ο αριθμός των ζευγών attributes για συσχέτιση είναι 14701753. Το ποσοστό των ζευγών attributes με συσχέτιση άνω του 0.8 είναι 31.16% και το ποσοστό ζευγών attributes με συσχέτιση ίση με 1 είναι 31.15%. Φυσικά, τα ζεύγη attributes που και τα δύο αναπαριστούν το ίδιο χαρακτηριστικό είναι υψηλώς συσχετισμένα, αλλά αυτή δεν είναι η μόνη περίπτωση. Σύμφωνα με τον πίνακα, θα έπρεπε να υπάρχουν πολύ λιγότερα χαρακτηριστικά, καθώς υπάρχουν πάρα πολλά που αντιστοιχούν στο ίδιο πράγμα και αποπροσανατολίζουν τον ταξινομητή.

5.3. Αποτελέσματα της Ταξινόμησης

Σύμφωνα με την διαδικασία ταξινόμησης και τα δεδομένα που αναφέρθηκαν προηγουμένως, υπάρχουν πολλά στοιχεία από τη συλλογή “Related to Food and Drink” που το προτεινόμενο σύστημα υπέδειξε πως θα έπρεπε να ανήκουν στο “Food and Drink Project”. Η πλειονότητα των στοιχείων που ταξινομήθηκαν ήταν αληθώς θετικά, πράγμα που υποδεικνύει την πολύ καλή απόδοση του ταξινομητή. Ακολούθως εμφανίζονται κάποια παραδείγματα των στοιχείων που ταξινομήθηκαν πως ανήκουν στο “Food and Drink Project”. Μια πλήρης συλλογή των ταξινομημένων στοιχείων βρίσκεται στην συλλογή “Classified Food and Drink” [22].

Traditional cooked English breakfast

The Wellcome Library

europeana.eu

Pieces of milk chocolate containing nuts

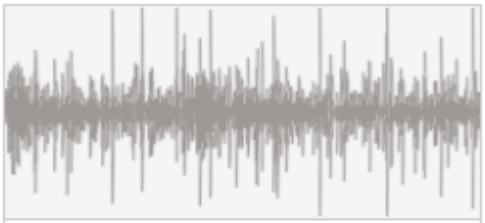
The Wellcome Library

europeana.eu

Bowl of cherries

The Wellcome Library

europeana.eu



Jeffries, Janis (2 of 10) National Life Stories Collection: Crafts' Lives

The British Library

europeana.eu

Το τελευταίο αρχείο audio με μία πρώτη ματιά φαίνεται πως είναι ταξινομημένο ψευδώς θετικά, όμως, παρατηρώντας την περιγραφή που συνοδεύει το αρχείο, γίνεται εμφανές πως άρμοζε η ταξινόμησή του στο "Food and Drink Project":

> *"… [0:06:50] Arguments and debates at the **dinner tables** of Jewish friends; comments from Harold Pinter about observing his family; lack of conversation in JJs home. [0:08:05] [Pause]. Being asked by students about going to **cocktail** parties. **Mealtimes** at home; having first **espresso** in Soho; discovering new **fruits** and **vegetables**; learning to **cook** from friends parents; chip fryer mentality. Parents responses to **food cooked** by JJ; **Bar** Italia in Soho; loathing for Camp **Coffee** and Carnation **Milk**; bringing home real **coffee** from Soho. Visiting markets while travelling; improvement in English **food** with the Common Market; **food cooked** by Jamaican friends from school; mothers knowledge of cuts of **meat**. …"*

5.4. Αποτελέσματα των Κανόνων Συσχέτισης

Οι κανόνες συσχέτισης που παράχθηκαν από το testing dataset, το οποίο αντιστοιχεί στην συλλογή “Related to Food and Drink” αρμόζουν εμφανώς στα δεδομένα. Αρχικά, η εμπιστοσύνη (confidence) έχει την υψηλότερη δυνατή τιμή για όλους τους κανόνες, ενώ παράχθηκαν 100 κανόνες, και επιπλέον, εύκολα παρατηρείται πως η σημασιολογία πίσω από τους κανόνες είναι συνεπής. Ακολούθως, παρουσιάζονται οι πρώτοι 10 κανόνες που παράχθηκαν.

Best rules found:

1.	CulturalObject_type_class = true	==>	
	hasContributor_inverseObjectProperty = false		conf:(1)
2.	CulturalObject_type_class = true	==>	
	hasCountry_inverseObjectProperty = false		conf:(1)
3.	hasCountry_objectProperty = true	==>	
	CulturalObject_type_class = true		conf:(1)
4.	CulturalObject_type_class = true	==>	
	hasCountry_objectProperty = true		conf:(1)
5.	CulturalObject_type_class = true	==>	
	hasCreator_inverseObjectProperty = false		conf:(1)
6.	CulturalObject_type_class = true	==>	
	hasLanguage_inverseObjectProperty = false		conf:(1)
7.	hasLanguage_objectProperty = true	==>	
	CulturalObject_type_class = true		conf:(1)
8.	CulturalObject_type_class = true	==>	
	hasLanguage_objectProperty = true		conf:(1)
9.	hasCountry_objectProperty = true	==>	
	hasContributor_inverseObjectProperty = false		conf:(1)
10.	hasLanguage_objectProperty = true	==>	
	hasContributor_inverseObjectProperty = false		conf:(1)

Κεφάλαιο 6. Συμπεράσματα και Μελλοντικές Εργασίες

Στο κεφάλαιο αυτό, εξάγονται συμπεράσματα σχετικά με το προτεινόμενο μοντέλο βασισμένα στα πειραματικά αποτελέσματα και στο τέλος παρατίθεται μία συζήτηση σχετικά με τις μελλοντικές επεκτάσεις του συστήματος.

6.1. Συμπεράσματα

Η παρούσα διπλωματική προτείνει ένα μοντέλο που εμπλουτίζει επιτυχώς τις οντολογίες αυτοματοποιημένα. Όσο είναι δυνατόν να γνωρίζουμε, τη συγκεκριμένη στιγμή, δεν υπάρχει κάποια σχετική υλοποίηση που να είναι ικανή να διαχειριστεί τις οντολογίες σαν σύνολα δεδομένων και να τις εμπλουτίσει με νέους ισχυρισμούς χωρίς ανθρώπινη βοήθεια. Όχι μόνο η διαδικασία ταξινόμησης, αλλά επίσης και η διαδικασία εκμάθησης κανόνων συσχέτισης έδειξε υποσχόμενα αποτελέσματα. Όπως είναι φυσικό, ο προκύπτων εμπλουτισμός εξαρτάται κυρίως από τα δεδομένα

εισόδου και από το κατά πόσο η πληροφορία που παρέχεται ως είσοδος είναι βοηθητική ή αποπροσανατολιστική για το πρόγραμμα και τις τεχνικές μηχανικής μάθησης που εφαρμόζονται. Με βάση τα δεδομένα που χρησιμοποιήθηκαν για την αξιολόγηση του συστήματος, μπορούμε να αποφανθούμε πως η χρήση του προτείνεται για παρόμοιες λειτουργίες ταξινόμησης δεδομένων και τα αποτελέσματα είναι τα επιθυμητά.

6.2. Μελλοντικές Εργασίες

Υπάρχει πληθώρα πιθανών επεκτάσεων του συστήματος που υλοποιήθηκε. Οι επεκτάσεις αυτές αφορούν την αύξηση των ικανοτήτων του συστήματος, όχι μόνο σχετικά με τις τεχνικές μηχανικής μάθησης που χρησιμοποιήθηκαν, αλλά επίσης σχετικά με την διαχείριση της πληροφορίας που αποκτάται, καθώς και από το βαθμό που ο χρήστης μπορεί να παρέμβει στα δεδομένα ώστε να παραχθεί μία ημιαυτόματη διαδικασία που θα παράγει πιο αξιόπιστα αποτελέσματα.

Όσον αφορά τις ικανότητες του συστήματος, οι επιλογές σχετικά με τους ταξινομητές και την εκμάθηση κανόνων συσχέτισης μπορούν εύκολα να αυξηθούν, όχι μόνο στο φάσμα των αλγορίθμων που παρέχει το Weka ως επιπρόσθετες υλοποιήσεις, αλλά επίσης με την εισαγωγή νέων αλγορίθμων απευθείας στο πρόγραμμα Java. Επιπλέον, δεν ήταν πάντοτε δυνατή η διαχείριση των κανόνων συσχέτισης που παρήχθησαν, λόγω της πολυπλοκότητας της συσχέτισης που απαιτείται για τον εμπλουτισμό της οντολογίας ή λόγω του τύπου των χαρακτηριστικών (attributes). Αυτό αποτελεί τροχοπέδη που θα μπορούσε εξίσου να αντιμετωπιστεί σε κάποια μελλοντική επέκταση.

Όσον αφορά τα δεδομένα εισόδου, μία μελλοντική επέκταση που πιθανότατα θα βελτίωνε σημαντικά τα αποτελέσματα είναι να εκτιμάται χειροκίνητα ο ταξινομητής και οι τύποι των χαρακτηριστικών και να συγχωνεύονται τα χαρακτηριστικά που είναι υψηλά συσχετιζόμενα. Όπως παρουσιάστηκε στο Κεφάλαιο 5, καθίσταται δυνατό δύο ή περισσότερα χαρακτηριστικά να αφορούν το ίδιο πράγμα και θα ήταν ωφέλιμο να συνδυάζονται με στόχο να αποφευχθεί ο αποπροσανατολισμός του ταξινομητή.

Αναφορές

[1] T. Berners-Lee, "Semantic Web Roadmap," 14 October 1998. [Online]. Available: https://www.w3.org/DesignIssues/Semantic.html.

[2] I. Herman, "W3C Technology and Society domain," October 2013. [Online]. Available: https://www.w3.org/2001/sw/Activity.

[3] T. R. Gruber, "Towards principles for the design of ontologies used for knowledge sharing," *International Journal of Human-Computer Studies,* vol. 43, no. 5-6, pp. 907-928, 1995.

[4] R. Struder, R. Benjamins and D. Fensel, "Knowledge engineering: Principles and methods," *Data & Knowledge Engineering,* vol. 25, no. 1-2, pp. 161-197, 1998.

[5] A. Artale, D. Calvanese, R. Kontchakov and M. Zakharyaschev, "The DL-lite family and relations," *Journal of Artificial Intelligence Research,* vol. 36, no. 1, pp. 1-69, 2009.

[6] P. Hitzler, M. Krotzsch, B. Parsia, P. Patel-Schneider and S. Rudolph, "OWL 2 Web Ontology Language Primer (Second Edition)," 11 December 2012. [Online]. Available: https://www.w3.org/TR/owl-primer/.

[7] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[8] D. Aha and D. Kibler, "Instance-based learning algorithms," *Machine learning,* vol. 6, pp. 37-66, 1991.

[9] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in *20th International Conference on Very Large Data Bases*, Los Altos, CA, 1994.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The Weka Data Mining Software: An Update," *SIGKDD Explorations,* vol. 11, no. 1, 2009.

[11] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM,* vol. 38, no. 11, pp. 39-41, 1995.

[12] P. Barrio, G. Simoes, H. Garhardas and L. Gravano, "REEL: A Relation Extraction Learning Framework," in *JCDL*, London, 2014.

[13] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka and T. M. Mitchell, "Toward an Architecture for Never-Ending Language Learning," in *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2010.

[14] F. M. Suchanek, M. Sozio and G. Weikum, "SOFIE: a self-organizing framework for information extraction," in *Proceedings of the 18th international conference on World wide web*, Madrid, 2009.

[15] S. Bloehdorn, C. Philipp, H. Andreas and S. Steffen, "An Ontology-based Framework for Text Mining," *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology,* vol. 20, no. 1, pp. 87-112, 2005.

[16] "Sample Ontology: Wine," [Online]. Available: http://protege.cim3.net/file/pub/ontologies/wine/wine.owl.

[17] "WITH," [Online]. Available: http://with.image.ntua.gr/assets/index.html#home.

[18] "F&D TopFoto Collection," [Online]. Available: http://with.image.ntua.gr/assets/index.html#collectionview/56ec6b0c75fe241fb97dc87a/count/40.

[19] "TopFoto EuropeanaPhotography Collection," [Online]. Available: http://with.image.ntua.gr/assets/index.html#collectionview/5746bf4d4c74792acafd6d1c/count/20.

[20] "Europeana Collections," [Online]. Available: http://www.europeana.eu/portal/.

[21] "Related to Food and Drink Collection," [Online]. Available: http://with.image.ntua.gr/assets/index.html#collectionview/56ed779b75fe2408ecdcc09d/count/20.

[22] "Classified Food and Drink Collection," [Online]. Available: http://with.image.ntua.gr/assets/index.html#collectionview/57713ec5713f2118278aaebf/count/220.